AEA Continuing Education Program

Cross-section Econometrics

Guido Imbens, Harvard University

January 8-10, 2012

AEA Lectures **Chicago, IL, January 2012**

Lecture 1, Sunday, Jan 7th,pm-pm

Estimation of Average Treatment Effects Under Unconfoundedness

1. Introduction

In this lecture we look at several methods for estimating average effects of a program, treatment, or regime, under unconfoundedness. The setting is one with a binary program. The traditional example in economics is that of a labor market program where some individuals receive training and others do not, and interest is in some measure of the effectiveness of the training. Unconfoundedness, a term coined by Rubin (1990), refers to the case where (non-parametrically) adjusting for differences in a fixed set of covariates removes biases in comparisons between treated and control units, thus allowing for a causal interpretation of those adjusted differences. This is perhaps the most important special case for estimating average treatment effects in practice. Alternatives typically involves strong assumptions linking unobservables to observables in specific ways in order to allow adjusting for the relevant differences in unobserved variables. An example of such a strategy is instrumental variables, which will be discussed in Lecture 3. A second example that does not involve additional assumptions is the bounds approach developed by Manski (1990, 2003).

Under the specific assumptions we make in this setting, the population average treatment effect can be estimated at the standard parametric $\sqrt{N}$ rate without functional form assumptions. A variety of estimators, at first sight quite different, have been proposed for implementing this. The estimators include regression estimators, propensity score based estimators and matching estimators. Many of these are used in practice, although rarely is this choice motivated by principled arguments. In practice the differences between the estimators are relatively minor when applied appropriately, although matching in combination with regression is generally more robust and is probably the recommended choice. More important than the choice of estimator are two other issues. Both involve analyses of the data without the outcome variable. First, one should carefully check the extent of the overlap

in covariate distributions between the treatment and control groups. Often there is a need for some trimming based on the covariate values if the original sample is not well balanced. Without this, estimates of average treatment effects can be very sensitive to the choice of, and small changes in the implementation of, the estimators. In this part of the analysis the propensity score plays an important role. Second, it is useful to do some assessment of the appropriateness of the unconfoundedness assumption. Although this assumption is not directly testable, its plausibility can often be assessed using lagged values of the outcome as pseudo outcomes. Another issue is variance estimation. For matching estimators bootstrapping, although widely used, has been shown to be invalid. We discuss general methods for estimating the conditional variance that do not involve resampling.

In these notes we first set up the basic framework and state the critical assumptions in Section 2. In Section 3 we describe the leading estimators. In Section 4 we discuss variance estimation. In Section 5 we discuss assessing one of the critical assumptions, unconfoundedness. In Section 6 we discuss dealing with a major problem in practice, lack of overlap in the covariate distributions among treated and controls. In Section 7 we illustrate some of the methods using a well known data set in this literature, originally put together by Lalonde (1986).

In these notes we focus on estimation and inference for treatment effects. We do not discuss here a recent literature that has taken the next logical step in the evaluation literature, namely the optimal assignment of individuals to treatments based on limited (sample) information regarding the efficacy of the treatments. See Manski (2004, 2005, Dehejia (2004), Hirano and Porter (2005).

2. Framework

The modern set up in this literature is based on the potential outcome approach developed by Rubin (1974, 1977, 1978), which view causal effects as comparisons of potential outcomes defined on the same unit. In this section we lay out the basic framework.

2.1 Definitions

We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population. We postulate the existence for each unit of a pair of potential outcomes, $Y_i(0)$ for the outcome under the control treatment and $Y_i(1)$ for the outcome under the active treatment. In addition, each unit has a vector of characteristics, referred to as covariates, pretreatment variables or exogenous variables, and denoted by X_i.[1] It is important that these variables are not affected by the treatment. Often they take their values prior to the unit being exposed to the treatment, although this is not sufficient for the conditions they need to satisfy. Importantly, this vector of covariates can include lagged outcomes. Finally, each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment and $W_i = 1$ if unit i receives the active treatment. We therefore observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Distributions of (W_i, Y_i, X_i) refer to the distribution induced by the random sampling from the population.

Several additional pieces of notation will be useful in the remainder of these notes. First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

$$e(x) = \Pr(W_i = 1 | X_i = x) = \mathbb{E}[W_i | X_i = x].$$

Also, define, for $w \in \{0, 1\}$, the two conditional regression and variance functions:

$$\mu_w(x) = \mathbb{E}[Y_i(w) | X_i = x], \qquad \sigma_w^2(x) = \mathbb{V}(Y_i(w) | X_i = x).$$

2.2 Estimands: Average Treatment Effects

[1] Calling such variables exogenous is somewhat at odds with several formal definitions of exogeneity (e.g., Engle, Hendry and Richard, 1974), as knowledge of their distribution can be informative about the average treatment effects. It does, however, agree with common usage. See for example, Manski, Sandefur, McLanahan, and Powers (1992, p. 28).

In this discussion we will primarily focus on a number of average treatment effects (ATEs). For a discussion of testing for the presence of any treatment effects under unconfoundedness see Crump, Hotz, Imbens and Mitnik (2007). Focusing on average effects is less limiting than it may seem, however, as this includes averages of arbitrary transformations of the original outcomes.[2] The first estimand, and the most commonly studied in the econometric literature, is the population average treatment effect (PATE):

$$\tau_P = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Alternatively we may be interested in the population average treatment effect for the treated (PATT, e.g., Rubin, 1977; Heckman and Robb, 1984):

$$\tau_{P,T} = \mathbb{E}[Y_i(1) - Y_i(0)|W = 1].$$

Most of the discussion in these notes will focus on τ_P, with extensions to $\tau_{P,T}$ available in the references.

We will also look at sample average versions of these two population measures. These estimands focus on the average of the treatment effect in the specific sample, rather than in the population at large. These include, the sample average treatment effect (SATE) and the sample average treatment effect for the treated (SATT):

$$\tau_S = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big), \qquad \text{and} \quad \tau_{S,T} = \frac{1}{N_T}\sum_{i:W_i=1}\Big(Y_i(1) - Y_i(0)\Big),$$

where $N_T = \sum_{i=1}^{N} W_i$ is the number of treated units. The sample average treatment effects have received little attention in the recent econometric literature, although it has a long tradition in the analysis of randomized experiments (e.g., Neyman, 1923). Without further assumptions, the sample contains no information about the population ATE beyond the

[2]Lehman (1974) and Doksum (1974) introduce quantile treatment effects as the difference in quantiles between the two marginal treated and control outcome distributions. Bitler, Gelbach and Hoynes (2002) estimate these in a randomized evaluation of a social program. Firpo (2003) develops an estimator for such quantiles under unconfoundedness.

sample ATE. To see this, consider the case where we observe the sample $(Y_i(0), Y_i(1), W_i, X_i)$, $i = 1, \ldots, N$; that is, we observe for each unit both potential outcomes. In that case the sample average treatment effect, $\tau_S = \sum_i (Y_i(1) - Y_i(0))/N$, can be estimated without error. Obviously the best estimator for the population average effect, τ_P, is τ_S. However, we cannot estimate τ_P without error even with a sample where all potential outcomes are observed, because we lack the potential outcomes for those population members not included in the sample. This simple argument has two implications. First, one can estimate the sample ATE at least as accurately as the population ATE, and typically more so. In fact, the difference between the two variances is the variance of the treatment effect, which is zero only when the treatment effect is constant. Second, a good estimator for one average treatment effect is automatically a good estimator for the other. One can therefore interpret many of the estimators for PATE or PATT as estimators for SATE or SATT, with lower implied standard errors.

The difference in asymptotic variances forces the researcher to take a stance on what the quantity of interest is. For example, in a specific application one can legitimately reach the conclusion that there is no evidence, at the 95% level, that the PATE is different from zero, whereas there may be compelling evidence that the SATE is positive. Typically researchers in econometrics have focused on the PATE, but one can argue that it is of interest, when one cannot ascertain the sign of the population-level effect, to know whether one can determine the sign of the effect for the sample. Especially in cases, which are all too common, where it is not clear whether the sample is representative of the population of interest, results for the sample at hand may be of considerable interest.

2.2 Identification

We make the following key assumption about the treatment assignment:

Assumption 1 (Unconfoundedness)

$$\Big(Y_i(0), Y_i(1)\Big) \;\perp\!\!\!\perp\; W_i \;\Big|\; X_i.$$

This assumption was first articulated in this form in Rosenbaum and Rubin (1983a). Lechner (1999, 2002) refers to this as the "conditional independence assumption," Following a parametric version of this in Heckman and Robb (1984) it is also referred to as "selection on observables." In the missing data literature the equivalent assumption is referred to as "missing at random."

To see the link with standard exogeneity assumptions, suppose that the treatment effect is constant: $\tau = Y_i(1) - Y_i(0)$ for all i. Suppose also that the control outcome is linear in X_i:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. Then we can write

$$Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i.$$

Given the constant treatment effect assumption, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous. Without this constant treatment effect assumption, however, unconfoundedness does not imply a linear relation with (mean-)independent errors.

Next, we make a second assumption regarding the joint distribution of treatments and covariates:

Assumption 2 (Overlap)

$$0 < \Pr(W_i = 1|X_i) < 1.$$

Rosenbaum and Rubin (1983a) refer to the combination of the two assumptions as "stongly ignorable treatment assignment." For many of the formal results one will also need smoothness assumptions on the conditional regression functions and the propensity score ($\mu_w(x)$ and $e(x)$), and moment conditions on $Y_i(w)$. I will not discuss these regularity conditions here. Details can be found in the references for the specific estimators given below.

There has been some controversy about the plausibility of Assumptions 1 and 2 in economic settings and thus the relevance of the econometric literature that focuses on estimation and inference under these conditions for empirical work. In this debate it has been argued that agents' optimizing behavior precludes their choices being independent of the potential outcomes, whether or not conditional on covariates. This seems an unduly narrow view. In response I will offer three arguments for considering these assumptions. The first is a statistical, data descriptive motivation. A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units. A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment). Such an analysis may not lead to the final word on the efficacy of the treatment, but the absence of such an analysis would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not. The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated. Economic theory can help in classifying variables into those that need to be adjusted for versus those that do not, on the basis of their role in the decision process (e.g., whether they enter the utility function or the constraints). Given that, the unconfoundedness assumption merely asserts that all variables that need to be adjusted for are observed by the researcher. This is an empirical question, and not one that should be controversial as a general principle. It is clear that settings where some of these covariates are not observed will require strong assumptions to allow for identification. Such assumptions include instrumental variables settings where some covariates are assumed to be independent of the potential outcomes. Absent those assumptions, typically only bounds can be identified (e.g., Manski, 1990, 1995).

A third, related, argument is that even when agents optimally choose their treatment, two agents with the same values for observed characteristics may differ in their treatment choices

without invalidating the unconfoundedness assumption if the difference in their choices is driven by differencese in unobserved characteristics that are themselves unrelated to the outcomes of interest. The plausability of this will depend critically on the exact nature of the optimization process faced by the agents. In particular it may be important that the objective of the decision maker is distinct from the outcome that is of interest to the evaluator. For example, suppose we are interested in estimating the average effect of a binary input (e.g., a new technology) on a firm's output. Assume production is a stochastic function of this input because other inputs (e.g., weather) are not under the firm's control, or $Y_i = g(W, \varepsilon_i)$. Suppose that profits are output minus costs, $\pi_i(w) = g(w, \varepsilon_i) - c_i \cdot w$, and also that a firm chooses a production level to maximize expected profits, equal to output minus costs:

$$W_i = \arg\max_w \mathbb{E}[\pi_i(w)|c_i] = \arg\max_w \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon_i) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of production given inputs, then unconfoundedness will hold as

$$(g(0, \varepsilon_i), g(1, \varepsilon_i)) \perp\!\!\!\perp c_i.$$

Note that under the same assumptions one cannot necessarily identify the effect of the input on profits since $(\pi_i(0), \pi_i(1))$ are not independent of c_i. See for a related discussion, in the context of instrumental variables, Athey and Stern (1998). Heckman, Lalonde and Smith (2000) discuss alternative models that justify unconfoundedness. In these models individuals do attempt to optimize the same outcome that is the variable of interest to the evaluator. They show that selection on observables assumptions can be justified by imposing restrictions

on the way individuals form their expectations about the unknown potential outcomes. In general, therefore, a researcher may wish to, either as a final analysis or as part of a larger investigation, consider estimates based on the unconfoundedness assumption.

Given strongly ignorable treatment assignment one can identify the population average treatment effect. The key insight is that given unconfoundedness, the following equalities holds:

$$\mu_w(x) = \mathbb{E}[Y_i(w)|X_i = x] = \mathbb{E}[Y_i(w)|W_i = w, X_i = x] = \mathbb{E}[Y_i|W_i = w, X_i = x],$$

and $\mu_w(x)$ is identified. Thus one can estimate the average treatment effect τ by first estimating the average treatment effect for a subpopulation with covariates $X = x$:

$$\begin{aligned}\tau(x) &\equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x] \\ &= \mathbb{E}[Y_i(1)|X_i = x, W_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, W_i = 0] \\ &= \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0].\end{aligned}$$

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y_i|X_i = x, W_i = w]$ for all values of w and x in the support of these variables. This is where the second assumption enters. If the overlap assumption is violated at $X = x$, it would be infeasible to estimate both $\mathbb{E}[Y_i|X_i = x, W_i = 1]$ and $\mathbb{E}[Y_i|X_i = x, W_i = 0]$ because at those values of x there would be either only treated or only control units.

Some researchers use weaker versions of the unconfoundedness assumption (e.g., Heckman, Ichimura, and Todd, 1998). If the interest is in the population average treatment effect, it is in fact sufficient to assume that

$$\mathbb{E}[Y_i(w)|W_i, X_i] = \mathbb{E}[Y_i(w)|X_i],$$

for $w = 0, 1$. Although this assumption is unquestionably weaker, in practice it is rare that a convincing case is made for the weaker assumption without the case being equally strong

for the stronger Assumption 1. The reason is that the weaker assumption is intrinsically tied to functional form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (e.g., logarithms) without the strong assumption.

One can weaken the unconfoundedness assumption in a different direction if one is only interested in the average effect for the treated (e.g., Heckman, Ichimura and Todd, 1997). In that case one need only assume $Y_i(0) \perp\!\!\!\perp W_i \,\Big|\, X_i$. and the weaker overlap assumption $\Pr(W_i = 1|X_i) < 1$. These two assumptions are sufficient for identification of PATT because moments of the distribution of $Y(1)$ for the treated are directly estimable.

An important result building on the unconfoundedness assumption shows that one need not condition simultaneously on all covariates. The following result shows that all biases due to observable covariates can be removed by conditioning solely on the propensity score:

Result 1 *Suppose that Assumption 1 holds. Then:*

$$\Big(Y_i(0), Y_i(1)\Big) \perp\!\!\!\perp W_i \,\Big|\, e(X_i).$$

Proof: We will show that $\Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) = \Pr(W_i = 1|e(X_i)) = e(X_i)$, implying independence of $(Y_i(0), Y_i(1))$ and W_i conditional on $e(X_i)$. First, note that

$$\begin{aligned}
\Pr(W_i = 1|Y_i(0), Y_i(1), e(X_i)) &= \mathbb{E}[W_i = 1|Y_i(0), Y_i(1), e(X_i)] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), e(X), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|Y_i(0), Y_i(1), X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}[W_i|X_i]\Big|Y_i(0), Y_i(1), e(X_i)\right] = \mathbb{E}\left[e(X_i)|Y_i(0), Y_i(1), e(X_i)\right] = e(X_i),
\end{aligned}$$

where the last equality but one follows from unconfoundedness. The same argument shows that

$$\Pr(W_i = 1|e(X_i)) = \mathbb{E}[W_i = 1|e(X_i)] = \mathbb{E}\left[\mathbb{E}[W_i = 1|X_i]\Big|e(X_i)\right] = \mathbb{E}\left[e(X_i)|e(X_i)\right] = e(X_i).$$

□

Extensions of this result to the multivalued treatment case are given in Imbens (2000) and Lechner (2001).

To provide intuition for the Rosenbaum-Rubin result, recall the textbook formula for omitted variable bias in the linear regression model. Suppose we have a regression model with two regressors:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i.$$

The bias of omitting X_i from the regression on the coefficient on W_i is equal to $\beta_2'\delta$, where δ is the vector of coefficients on W_i in regressions of the elements of X_i on W_i. By conditioning on the propensity score we remove the correlation between X_i and W_i because $X_i \perp\!\!\!\perp W_i | e(X_i)$. Hence omitting X_i no longer leads to any bias (although it may still lead to some efficiency loss).

2.4 Efficiency Bounds and Asymptotic Variances for Population Average Treatment Effects

Next we review some results on the effiency bound for estimators of the average treatment effects τ_P. This requires strong ignorability and some smoothness assumptions on the conditional expectations of potential outcomes and the treatment indicator (for details, see Hahn, 1998). Formally, Hahn (1998) shows that for any regular estimator for τ_P, denoted by $\hat{\tau}$, with

$$\sqrt{N} \cdot (\hat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}(0, V),$$

we can show that

$$V \geq \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\tau(X_i) - \tau_P)^2 \right]. \tag{1}$$

Knowing the propensity score does not affect this efficiency bound.

Hahn also shows that asymptotically linear estimators exist that achieve the efficiency bound, and hence such efficient estimators can be approximated as

$$\hat{\tau} = \tau_P + \frac{1}{N}\sum_{i=1}^{N} \psi(Y_i, W_i, X_i, \tau_P) + o_p(N^{-1/2}),$$

where $\psi(\cdot)$ is the efficient score:

$$\psi(y, w, x, \tau_P) = \left(\frac{wy}{e(x)} - \frac{(1-w)y}{1-e(x)}\right) - \tau_P - \left(\frac{\mu_1(x)}{e(x)} + \frac{\mu_0(x)}{1-e(x)}\right)\cdot(w - e(x)). \tag{2}$$

3. Estimating Average Treatment Effects

Here we discuss the leading estimators for average treatment effects under unconfoundedness. What is remarkable about this literature is the wide range of ostensibly quite different estimators, many of which are regularly used in empirical work. We first briefly describe a number of the estimators, and then discuss their relative merits.

3.1 Regression

The first class of estimators relies on consistent estimation of $\mu_w(x)$ for $w = 0, 1$. Given $\hat{\mu}_w(x)$ for these regression functions, the PATE and SATE are estimated by averaging their difference over the empirical distribution of the covariates:

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\Big). \tag{3}$$

In most implementations the average of the predicted treated outcome for the treated is equal to the average observed outcome for the treated (so that $\sum_i W_i \cdot \hat{\mu}_1(X_i) = \sum_i W_i \cdot Y_i$), and similarly for the controls, implying that $\hat{\tau}_{\text{reg}}$ can also be written as

$$\hat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} W_i \cdot \Big(Y_i - \hat{\mu}_0(X_i)\Big) + (1 - W_i) \cdot \Big(\hat{\mu}_1(X_i) - Y_i\Big).$$

Early estimators for $\mu_w(x)$ included parametric regression functions, for example linear regression (e.g., Rubin, 1977). Such parametric alternatives include least squares estimators

with the regression function specified as

$$\mu_w(x) = \beta' x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ. In this case one can estimate τ simply by least squares estimation using the regression function

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i.$$

More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$. In that case one estimate the two regression functions separately on the two subsamples and then substitute the predicted values in (3).

These simple regression estimators can be sensitive to differences in the covariate distributions for treated and control units. The reason is that in that case the regression estimators rely heavily on extrapolation. To see this, note that the regression function for the controls, $\mu_0(x)$ is used to predict missing outcomes for the treated. Hence on average one wishes to use predict the control outcome at $\overline{X}_T = \sum_i W_i \cdot X_i / N_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \hat{\beta}'(\overline{X}_T - \overline{X}_C)$. If $\overline{X}_T$ and the average covariate value for the controls, $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. However, with the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.

More recently, nonparametric estimators have been proposed. Imbens, Newey and Ridder (2005) and Chen, Hong, and Tarozzi (2005) propose estimating $\mu_w(x)$ through series or sieve methods. A simple version of that with a scalar X would specify the regression function as

$$\mu_w(x) = \sum_{l=0}^{L_N} \beta_{w,l} \cdot x^k,$$

with L_N, the number of terms in the polynomial expansion, an increasing function of the sample size. They show that this estimator for τ_P achieves the semiparametric efficiency

bounds. Heckman, Ichimura and Todd (1997, 1998), and Heckman, Ichimura, Smith and Todd (1998) consider kernel methods for estimating $\mu_w(x)$, in particular focusing on local linear approaches. Given a kernel $K(\cdot)$, and a bandwidth h_N let

$$\left(\hat{\alpha}_{w,x},\hat{\beta}_{w,x}\right) = \arg\min_{\alpha_{w,x},\beta_{w,x}} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h_N}\right) \cdot \left(Y_i - \alpha_{w,x} - \beta_{w,x}\cdot X_i\right)^2,$$

leading to the estimator

$$\hat{\mu}_w(x) = \hat{\alpha}_{w,x}.$$

3.2 Matching

Regression estimators impute the missing potential outcomes using the estimated regression function. Thus, if $W_i = 1$, $Y_i(1)$ is observed and $Y_i(0)$ is missing and imputed with a consistent estimator $\hat{\mu}_0(X_i)$ for the conditional expectation. Matching estimators also impute the missing potential outcomes, but do so using only the outcomes of nearest neighbours of the opposite treatment group. In that sense matching is similar to nonparametric kernel regression methods, with the number of neighbors playing the role of the bandwidth in the kernel regression. In fact, matching can be inrrepreted as a limiting version of the standard kernel estimator where the bandwidth goes to zero. This minimizes the bias among nonnegative kernels, but potentially increases the variance relative to kernel estimators. A formal difference with kernel estimators is that the asymptotic distribution is derived conditional on the implicit bandwidth, that is, the number of neighbours, which is often fixed at one. Using such asymptotics, the implicit estimate $\hat{\mu}_w(x)$ is (close to) unbiased, but not consistent for $\mu_w(x)$. In contrast, the regression estimators discussed earlier relied on the consistency of $\mu_w(x)$.

Matching estimators have the attractive feature that given the matching metric, the researcher only has to choose the number of matches. In contrast, for the regression estimators discussed above, the researcher must choose smoothing parameters that are more difficult to interpret; either the number of terms in a series or the bandwidth in kernel regression.

Within the class of matching estimators, using only a single match leads to the most credible inference with the least bias, at most sacrificing some precision. This can make the matching estimator easier to use than those estimators that require more complex choices of smoothing parameters, and may explain some of its popularity.

Matching estimators have been widely studied in practice and theory (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1989, 1995, 2002; Rubin, 1973b, 1979; Heckman, Ichimura and Todd, 1998; Dehejia and Wahba, 1999; Abadie and Imbens, 2002, AI). Most often they have been applied in settings with the following two characteristics: (i) the interest is in the average treatment effect for the treated, and (ii), there is a large reservoir of potential controls. This allows the researcher to match each treated unit to one or more distinct controls (referred to as matching without replacement). Given the matched pairs, the treatment effect within a pair is then estimated as the difference in outcomes, with an estimator for the PATT obtained by averaging these within-pair differences. Since the estimator is essentially the difference in two sample means, the variance is calculated using standard methods for differences in means or methods for paired randomized experiments. The remaining bias is typically ignored in these studies. The literature has studied fast algorithms for matching the units, as fully efficient matching methods are computationally cumbersome (e.g., Gu and Rosenbaum, 1993; Rosenbaum, 1995). Note that in such matching schemes the order in which the units are matched is potentially important.

Here we focus on matching estimators for PATE and SATE. In order to estimate these targes we need to match both treated and controls, and allow for matching with replacement. Formally, given a sample, $\{(Y_i, X_i, W_i)\}_{i=1}^N$, let $\ell_m(i)$ be the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\Big\{\|X_j - X_i\| \leq \|X_l - X_i\|\Big\} = m,$$

where $1\{\cdot\}$ is the indicator function, equal to one if the expression in brackets is true and zero otherwise. In other words, $\ell_m(i)$ is the index of the unit in the opposite treatment group that is the m-th closest to unit i in terms of the distance measure based on the norm $\|\cdot\|$.

In particular, $\ell_1(i)$ is the nearest match for unit i. Let $\mathcal{J}_M(i)$ denote the set of indices for the first M matches for unit i: $\mathcal{J}_M(i) = \{\ell_1(i), \ldots, \ell_M(i)\}$. Define the imputed potential outcomes as:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases} \qquad \hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

The simple matching estimator is then

$$\hat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{Y}_i(1) - \hat{Y}_i(0)\Big). \tag{4}$$

AI show that the bias of this estimator is of order $O(N^{-1/K})$, where K is the dimension of the covariates. Hence, if one studies the asymptotic distribution of the estimator by normalizing by $\sqrt{N}$ (as can be justified by the fact that the variance of the estimator is of order $O(1/N)$), the bias does not disappear if the dimension of the covariates is equal to two, and will dominate the large sample variance if K is at least three.

Let us make clear three caveats to the AI result. First, it is only the continuous covariates that should be counted in K. With discrete covariates the matching will be exact in large samples, therefore such covariates do not contribute to the order of the bias. Second, if one matches only the treated, and the number of potential controls is much larger than the number of treated units, one can justify ignoring the bias by appealling to an asymptotic sequence where the number of potential controls increases faster than the number of treated units. Specifically, if the number of controls, N_0, and the number of treated, N_1, satisfy $N_1/N_0^{4/K} \to 0$, then the bias disappears in large samples after normalization by $\sqrt{N}_1$. Third, even though the order of the bias may be high, the actual bias may still be small if the coefficients in the leading term are small. This is possible if the biases for different units are at least partially offsetting. For example, the leading term in the bias relies on the regression function being nonlinear, and the density of the covariates having a nonzero slope. If either the regression function is close to linear, or the density of the covariates close to constant, the resulting bias may be fairly limited. To remove the bias, AI suggest combining the matching process with a regression adjustment.

Another point made by AI is that matching estimators are generally not efficient. Even in the case where the bias is of low enough order to be dominated by the variance, the estimators are not efficient given a fixed number of matches. To reach efficiency one would need to increase the number of matches with the sample size, as done implicitly in kernel estimators. In practice the efficiency loss is limited though, with the gain of going from two matches to a large number of matches bounded as a fraction of the standard error by 0.16 (see AI).

In the above discussion the distance metric in choosing the optimal matches was the standard Euclidan metric $d_E(x,z) = (x-z)'(x-z)$. All of the distance metrics used in practice standardize the covariates in some manner. The most popular metrics are the Mahalanobis metric, where

$$d_M(x,z) = (x-z)'(\Sigma_X^{-1})(x-z),$$

where Σ is covariance matrix of the covairates, and the diagonal version of that

$$d_{AI}(x,z) = (x-z)'\text{diag}(\Sigma_X^{-1})(x-z).$$

Note that depending on the correlation structure, using the Mahalanobis metric can lead to situations where a unit with $X_i = (5,5)$ is a closer match for a unith with $X_i = (0,0)$ than a unit with $X_i = (1,4)$, despite being further away in terms of each covariate separately.

3.3 Propensity Score Methods

Since the work by Rosenbaum and Rubin (1983a) there has been considerable interest in methods that avoid adjusting directly for all covariates, and instead focus on adjusting for differences in the propensity score, the conditional probability of receiving the treatment. This can be implemented in a number of different ways. One can weight the observations in terms of the propensity score (and indirectly also in terms of the covariates) to create balance between treated and control units in the weighted sample. Hirano, Imbens and Ridder (2003) show how such estimators can achieve the semiparametric efficiency bound.

Alternatively one can divide the sample into subsamples with approximately the same value of the propensity score, a technique known as blocking. Finally, one can directly use the propensity score as a regressor in a regression approach or match on the propensity score.

If the researcher knows the propensity score all three of these methods are likely to be effective in eliminating bias. Even if the resulting estimator is not fully efficient, one can easily modify it by using a parametric estimate of the propensity score to capture most of the efficiency loss. Furthermore, since these estimators do not rely on high-dimensional nonparametric regression, this suggests that their finite sample properties would be attractive.

In practice the propensity score is rarely known, and in that case the advantages of the estimators discussed below are less clear. Although they avoid the high-dimensional nonparametric estimation of the two conditional expectations $\mu_w(x)$, they require instead the equally high-dimensional nonparametric estimation of the propensity score. In practice the relative merits of these estimators will depend on whether the propensity score is more or less smooth than the regression functions, or whether additional information is available about either the propensity score or the regression functions.

3.3.1 Weighting

The first set of "propensity score" estimators use the propensity score as weights to create a balanced sample of treated and control observations. Simply taking the difference in average outcomes for treated and controls,

$$\hat{\tau} = \frac{\sum W_i Y_i}{\sum W_i} - \frac{\sum (1-W_i) Y_i}{\sum 1 - W_i},$$

is not unbiased for $\tau^P = \mathbb{E}[Y_i(1) - Y_i(0)]$ because, conditional on the treatment indicator, the distributions of the covariates differ. By weighting the units by the inverse of the probability of receiving the treatment, one can undo this imbalance. Formally, weighting estimators rely on the equalities:

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\frac{WY_i(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY_i(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e(X)Y_i(1)}{e(X)}\right]\right] = \mathbb{E}[Y_i(1)],$$

and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y_i(0)],$$

implying

$$\tau_P = \mathbb{E}\left[\frac{W\cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right). \tag{5}$$

In this particular form this is not necessarily an attractive estimator. The main reason is that, although the estimator can be written as the difference between a weighted average of the outcomes for the treated units and a weighted average of the outcomes for the controls, the weights do not necessarily add to one. Specifically, in (5), the weights for the treated units add up to $(\sum W_i/e(X_i))/N$. In expectation this is equal to one, but since its variance is positive, in any given sample some of the weights are likely to deviate from one. One approach for improving this estimator is simply to normalize the weights to unity. One can further normalize the weights to unity within subpopulations as defined by the covariates. In the limit this leads to the estimator proposed by Hirano, Imbens and Ridder (2003) who suggest using a nonparametric series estimator for $e(x)$. More precisely, they first specify a sequence of functions of the covariates, e.g., a power series, $h_l(x)$, $l = 1,\ldots,\infty$. Next, they choose a number of terms, $L(N)$, as a function of the sample size, and then estimate the L-dimensional vector γ_L in

$$\Pr(W=1|X=x) = \frac{\exp((h_1(x),\ldots,h_L(x))\gamma_L)}{1+\exp((h_1(x),\ldots,h_L(x))\gamma_L)},$$

by maximizing the associated likelihood function. Let $\hat{\gamma}_L$ be the maximum likelihood estimate. In the third step, the estimated propensity score is calculated as:

$$\hat{e}(x) = \frac{\exp((h_1(x),\ldots,h_L(x))\hat{\gamma}_L)}{1+\exp((h_1(x),\ldots,h_L(x))\hat{\gamma}_L)}.$$

Finally they estimate the average treatment effect as:

$$\hat{\tau}_{\text{weight}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\hat{e}(X_i)} - \sum_{i=1}^{N} \frac{(1-W_i) \cdot Y_i}{1-\hat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{(1-W_i)}{1-\hat{e}(X_i)}. \tag{6}$$

Hirano, Imbens and Ridder (2003) show that this estimator is efficient, whereas with the true propensity score the estimator would not be fully efficient (and in fact not very attractive).

This estimator highlights one of the interesting features of the problem of efficiently estimating average treatment effects. One solution is to estimate the two regression functions $\mu_w(x)$ nonparametrically; that solution completely ignores the propensity score. A second approach is to estimate the propensity score nonparametrically, ignoring entirely the two regression functions. If appropriately implemented, both approaches lead to fully efficient estimators, but clearly their finite sample properties may be very different, depending, for example, on the smoothness of the regression functions versus the smoothness of the propensity score. If there is only a single binary covariate, or more generally with only discrete covariates, the weighting approach with a fully nonparametric estimator for the propensity score is numerically identical to the regression approach with a fully nonparametric estimator for the two regression functions.

One difficulty with the weighting estimators that are based on the estimated propensity score is again the problem of choosing the smoothing parameters. Hirano, Imbens and Ridder (2003) use series estimators, which requires choosing the number of terms in the series. Ichimura and Linton (2001) consider a kernel version, which involves choosing a bandwidth. Theirs is currently one of the few studies considering optimal choices for smoothing parameters that focuses specifically on estimating average treatment effects. A departure from standard problems in choosing smoothing parameters is that here one wants to use nonparametric regression methods even if the propensity score is known. For example, if the probability of treatment is constant, standard optimality results would suggest using a high degree of smoothing, as this would lead to the most accurate estimator for the propensity score. However, this would not necessarily lead to an efficient estimator for the average treatment effect of interest.

3.3.2 Blocking on the Propensity Score

In their original propensity score paper Rosenbaum and Rubin (1983a) suggest the following "blocking propensity score" estimator. Using the (estimated) propensity score, divide the sample into M blocks of units of approximately equal probability of treatment, letting J_{im} be an indicator for unit i being in block m. One way of implementing this is by dividing the unit interval into M blocks with boundary values equal to m/M for $m = 1, \ldots, M-1$, so that

$$J_{im} = 1\{(m-1)/M < e(X_i) \leq m/M\},$$

for $m = 1, \ldots, M$. Within each block there are N_{wm} observations with treatment equal to w, $N_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Given these subgroups, estimate within each block the average treatment effect as if random assignment holds,

$$\hat{\tau}_m = \frac{1}{N_{1m}} \sum_{i=1}^{N} J_{im} W_i Y_i - \frac{1}{N_{0m}} \sum_{i=1}^{N} J_{im}(1 - W_i) Y_i.$$

Then estimate the overall average treatment effect as:

$$\hat{\tau}_{\text{block}} = \sum_{m=1}^{M} \hat{\tau}_m \cdot \frac{N_{1m} + N_{0m}}{N}.$$

Blocking can be interpreted as a crude form of nonparametric regression where the unknown function is approximated by a step function with fixed jump points. To establish asymptotic properties for this estimator would require establishing conditions on the rate at which the number of blocks increases with the sample size. With the propensity score known, these are easy to determine; no formal results have been established for the unknown case.

The question arises how many blocks to use in practice. Cochran (1968) analyses a case with a single covariate, and, assuming normality, shows that using five blocks removes at least 95% of the bias associated with that covariate. Since all bias, under unconfoudnedness, is

associated with the propensity score, this suggests that under normality five blocks removes most of the bias associated with all the covariates. This has often been the starting point of empirical analyses using this estimator (e.g., Rosenbaum and Rubin, 1983b; Dehejia and Wahba, 1999), and has been implemented in STATA by Becker and Ichino (2002). Often, however, researchers subsequently check the balance of the covariates within each block. If the true propensity score per block is constant, the distribution of the covariates among the treated and controls should be identical, or, in the evaluation terminology, the covariates should be balanced. Hence one can assess the adequacy of the statistical model by comparing the distribution of the covariates among treated and controls within blocks. If the distributions are found to be different, one can either split the blocks into a number of subblocks, or generalize the specification of the propensity score. Often some informal version of the following algorithm is used: If within a block the propensity score itself is unbalanced, the blocks are too large and need to be split. If, conditional on the propensity score being balanced, the covariates are unbalanced, the specification of the propensity score is not adequate. In the illustrations in the next lecture a particular algorithm is decribed for choosing the blocks.

3.3.3 Regression on the Propensity Score

The third method of using the propensity score is to estimate the conditional expectation of Y given W and $e(X)$ and average the difference. Although this method has been used in practice, there is no particular reason why this is an attractive method compared to the regression methods based on the covariates directly. In addition, the large sample properties have not been established.

3.3.4 Matching on the Propensity Score

The Rosenbaum-Rubin result implies that it is sufficient to adjust solely for differences in the propensity score between treated and control units. Since one of the ways in which one can adjust for differences in covariates is matching, another natural way to use the propensity score is through matching. Because the propensity score is a scalar function of the covariates,

the bias results in Abadie and Imbens (2002) imply that the bias term is of lower order than the variance term and matching leads to a $\sqrt{N}$-consistent, asymptotically normally distributed estimator. The variance for the case with matching on the true propensity score also follows directly from their results. More complicated is the case with matching on the estimated propensity score. We are not aware of any results that give the asymptotic variance for this case.

3.4. Mixed Methods

A number of approaches have been proposed that combine two of the three methods described earlier, typically regression with one of its alternatives. These methods appear to be the most attractive in practice. The motivation for these combinations is that, although one method alone is often sufficient to obtain consistent or even efficient estimates, incorporating regression may eliminate remaining bias and improve precision. This is particularly useful because neither matching nor the propensity score methods directly address the correlation between the covariates and the outcome. The benefit associated with combining methods is made explicit in the notion developed by Robins and Ritov (1997) of "double robustness." They propose a combination of weighting and regression where, as long as the parametric model for either the propensity score or the regression functions is specified correctly, the resulting estimator for the average treatment effect is consistent. Similarly, because matching is consistent with few assumptions beyond strong ignorability, thus methods that combine matching and regressions are robust against misspecification of the regression function.

3.4.1 Weighting and Regression

One can rewrite the HIR weighting estimator discussed above as estimating the following regression function by weighted least squares,

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i,$$

with weights equal to

$$\lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}}.$$

Without the weights the least squares estimator would not be consistent for the average treatment effect; the weights ensure that the covariates are uncorrelated with the treatment indicator and hence the weighted estimator is consistent.

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example as

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i. Such an estimator, using a more general semiparametric regression model, is suggested in Robins and Rotnitzky (1995), Robins, Rotnitzky and Zhao (1995), Robins and Ritov (1997), and implemented in Hirano and Imbens (2001). In the parametric context Robins and Ritov argue that the estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are specified correctly. That is, in the Robins-Ritov terminology, the estimator is doubly robust.

3.4.2 Blocking and Regression

Rosenbaum and Rubin (1983b) suggest modifying the basic blocking estimator by using least squares regression within the blocks. Without the additional regression adjustment the estimated treatment effect within blocks can be written as a least squares estimator of τ_m for the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \varepsilon_i,$$

using only the units in block m. As above, one can also add covariates to the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \beta_m' X_i + \varepsilon_i,$$

again estimated on the units in block m.

3.4.3 Matching and Regression

Since Abadie and Imbens (2002) show that the bias of the simple matching estimator can dominate the variance if the dimension of the covariates is too large, additional bias corrections through regression can be particularly relevant in this case. A number of such corrections have been proposed, first by Rubin (1973b) and Quade (1982) in a parametric setting. Let $\hat{Y}_i(0)$ and $\hat{Y}_i(1)$ be the observed or imputed potential outcomes for unit i; where these estimated potential outcomes equal observed outcomes for some unit i and its match $\ell(i)$. The bias in their comparison, $\mathbb{E}[\hat{Y}_i(1) - \hat{Y}_i(0)] - (Y_i(1) - Y_i(0))$, arises from the fact that the covariates for units i and $\ell(i)$, X_i and $X_{\ell(i)}$ are not equal, although close because of the matching process.

To further explore this, focusing on the single match case, define for each unit:

$$\hat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)} & \text{if } W_i = 1, \end{cases} \qquad \hat{X}_i(1) = \begin{cases} X_{\ell(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

If the matching is exact $\hat{X}_i(0) = \hat{X}_i(1)$ for each unit. If not, these discrepancies will lead to potential bias. The difference $\hat{X}_i(1) - \hat{X}_i(0)$ will therefore be used to reduce the bias of the simple matching estimator.

Suppose unit i is a treated unit ($W_i = 1$), so that $\hat{Y}_i(1) = Y_i(1)$ and $\hat{Y}_i(0)$ is an imputed value for $Y_i(0)$. This imputed value is unbiased for $\mu_0(X_{\ell(i)})$ (since $\hat{Y}_i(0) = Y_{\ell(i)}$), but not necessarily for $\mu_0(X_i)$. One may therefore wish to adjust $\hat{Y}_i(0)$ by an estimate of $\mu_0(X_i) - \mu_0(X_{\ell(i)})$. Typically these corrections are taken to be linear in the difference in the covariates for units i and its match, that is, of the form $\beta_0'(\hat{X}_i(1) - \hat{X}_i(0) = \beta_0'(X_i - X_{\ell(i)})$. One proposed correction is to estimate $\mu_0(x)$ directly by taking the control units that are used as matches for the treated units, with weights corresponding to the number of times a control observations is used as a match, and estimate a linear regression of the form

$$Y_i = \alpha_0 + \beta_0' X_i + \varepsilon_i,$$

on the weighted control observations by least squares. (If unit i is a control unit the correction would be done using an estimator for the regression function $\mu_1(x)$ based on a linear specification $Y_i = \alpha_1 + \beta_1' X_i$ estimated on the treated units.) AI show that if this correction is done nonparametrically, the resulting matching estimator is consistent and asymptotically normal, with its bias dominated by the variance.

4. Estimating Variances

The variances of the estimators considered so far typically involve unknown functions. For example, as discussed earlier, the variance of efficient estimators of PATE is equal to

$$V_P = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \tau)^2\right],$$

involving the two regression functions, the two conditional variances and the propensity score.

4.1 Estimating The Variance of Efficient Estimators for τ_P

For efficient estimators for τ_P the asymptotic variance is equal to the efficiency bound V_P. There are a number of ways we can estimate this. The first is essentially by brute force. All five components of the variance, $\sigma_0^2(x)$, $\sigma_1^2(x)$, $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, are consistently estimable using kernel methods or series, and hence the asymptotic variance can be estimated consistently. However, if one estimates the average treatment effect using only the two regression functions, it is an additional burden to estimate the conditional variances and the propensity score in order to estimate V_P. Similarly, if one efficiently estimates the average treatment effect by weighting with the estimated propensity score, it is a considerable additional burden to estimate the first two moments of the conditional outcome distributions just to estimate the asymptotic variance.

A second method applies to the case where either the regression functions or the propensity score is estimated using series or sieves. In that case one can interpret the estimators, given the number of terms in the series, as parametric estimators, and calculate the variance this way. Under some conditions that will lead to valid standard errors and confidence

intervals.

A third approach is to use bootstrapping (Efron and Tibshirani, 1993; Horowitz, 2002). Although there is little formal evidence specific for these estimators, given that the estimators are asymptotically linear, it is likely that bootstrapping will lead to valid standard errors and confidence intervals at least for the regression and propensity score methods. Bootstrapping is not valid for matching estimators, as shown by Abadie and Imbens (2007) Subsampling (Politis and Romano, 1999) will still work in this setting.

4.2 Estimating The Conditional Variance

Here we focus on estimation of the variance of estimators for τ_S, which is the conditional variance of the various estimators, conditional on the covariates $\mathbf{X}$ and the treatment indicators $\mathbf{W}$. All estimators used in practice are linear combinations of the outcomes,

$$\hat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

with the $\lambda(\mathbf{X}, \mathbf{W})$ known functions of the covariates and treatment indicators. Hence the conditional variance is

$$V\left(\hat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma^2_{W_i}(X_i).$$

The only unknown component of this variance is $\sigma^2_w(x)$. Rather than estimating this through nonparametric regression, AI suggest using matching to estimate $\sigma^2_w(x)$. To estimate $\sigma^2_{W_i}(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $v(i)$ be the closest unit to i with the same treatment indicator ($W_{v(i)} = W_i$). The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

$$\hat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{v(i)}\right)^2 /2.$$

Note that this estimator is not consistent estimators of the conditional variances. However this is not important, as we are interested not in the variances at specific points in the

covariates distribution, but in the variance of the average treatment effect. Following the process introduce above, this is estimated as:

$$\hat{V}\left(\hat{\tau}|\mathbf{X},\mathbf{W}\right)=\sum_{i=1}^{N}\lambda_i(\mathbf{X},\mathbf{W})^2\cdot\hat{\sigma}^2_{W_i}(X_i).$$

5. Assessing Unconfoundedness

The unconfoundedness assumption used throughout this discussion is not directly testable. It states that the conditional distribution of the outcome under the control treatment, $Y_i(0)$, given receipt of the active treatment and given covariates, is identical to the distribution of the control outcome given receipt of the control treatment and given covariates. The same is assumed for the distribution of the active treatment outcome, $Y_i(1)$. Yet since the data are completely uninformative about the distribution of $Y_i(0)$ for those who received the active treatment and of $Y_i(1)$ for those receiving the control, the data cannot directly reject the unconfoundedness assumption. Nevertheless, there are often indirect ways of assessing the this, a number of which are developed in Heckman and Hotz (1989) and Rosenbaum (1987). These methods typically rely on estimating a pseudo causal effect that is known to equal zero. If based on a statistical test we reject the null hypothesis that this causal effect varies from zero, the unconfoundedness assumption is considered less plausible. These tests can be divided into two broad groups.

The first set of tests focuses on estimating the causal effect of a treatment that is known not to have an effect, relying on the presence of multiple control groups (Rosenbaum, 1987). Suppose one has two potential control groups, for example eligible nonparticipants and ineligibles, as in Heckman, Ichimura and Todd (1997). One interpretation of the test is to compare average treatment effects estimated using each of the control groups. This can also be interpreted as estimating an "average treatment effect" using only the two control groups, with the treatment indicator now a dummy for being a member of the first group. In that case the treatment effect is known to be zero, and statistical evidence of a non-zero effect implies that at least one of the control groups is invalid. Again, not rejecting the test does not imply the unconfoundedness assumption is valid (as both control groups could

suffer the same bias), but non-rejection in the case where the two control groups are likely to have different biases makes it more plausible that the unconfoundness assumption holds. The key for the power of this test is to have available control groups that are likely to have different biases, if at all. Comparing ineligibles and eligible nonparticipants is a particularly attractive comparison. Alternatively one may use different geographic controls, for example from areas bordering on different sides of the treatment group.

One can formalize this test by postulating a three-valued indicator $T_i \in \{-0, 1, 1\}$ for the groups (e.g., ineligibles, eligible nonnonparticipants and participants), with the treatment indicator equal to $W_i = 1\{T_i = 1\}$, so that

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i \in \{-1, 0\} \\ Y_i(1) & \text{if } T_i = 1. \end{cases}$$

If one extends the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp T_i \,\Big|\, X_i,$$

then a testable implication is

$$Y_i(0) \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\},$$

and thus

$$Y_i \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\}.$$

An implication of this independence condition is being tested by the tests discussed above. Whether this test has much bearing on the unconfoundedness assumption depends on whether the extension of the assumption is plausible given unconfoundedness itself.

The second set of tests of unconfoundedness focuses on estimating the causal effect of the treatment on a variable known to be unaffected by it, typically because its value is

determined prior to the treatment itself. Such a variable can be time-invariant, but the most interesting case is in considering the treatment effect on a lagged outcome, commonly observed in labor market programs. If the estimated effect differs from zero, this implies that the treated observations are different from the controls in terms of this particular covariate given the others. If the treatment effect is estimated to be close to zero, it is more plausible that the unconfoundedness assumption holds. Of course this does not directly test this assumption; in this setting, being able to reject the null of no effect does not directly reflect on the hypothesis of interest, unconfoundedness. Nevertheless, if the variables used in this proxy test are closely related to the outcome of interest, the test arguably has more power. For these tests it is clearly helpful to have a number of lagged outcomes.

To formalize this, let us suppose the covariates consist of a number of lagged outcomes $Y_{i,-1},\ldots,Y_{i,-T}$ as well as time-invariant individual characteristics Z_i, so that $X_i = (Y_{i,-1},\ldots,Y_{i,-T},Z_i)$. By construction only units in the treatment group after period -1 receive the treatment; all other observed outcomes are control outcomes. Also suppose that the two potential outcomes $Y_i(0)$ and $Y_i(1)$ correspond to outcomes in period zero. Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

$$Y_{i,0}(1), Y_{i,0}(0) \;\perp\!\!\!\perp\; W_i \;\Big|\; Y_{i,-1},\ldots,Y_{i,-(T-1)},Z_i,$$

and the second assumes stationarity and exchangeability:

$$f_{Y_{i,s}(0)|Y_{i,s-1}(0),\ldots,Y_{i,s-(T-1)}(0),Z_i,W_i}(y_s|y_{s-1},\ldots,y_{s-(T-1)},z,w), \quad \text{does not depend on } i \text{ and } s.$$

Then it follows that

$$Y_{i,-1} \;\perp\!\!\!\perp\; W_i \;\Big|\; Y_{i,-2},\ldots,Y_{i,-T},Z_i,$$

which is testable. This hypothesis is what the procedure described above tests. Whether this test has much bearing on unconfoundedness depends on the link between the two assumptions and the original unconfoundedness assumption. With a sufficient number of lags

unconfoundedness given all lags but one appears plausible conditional on unconfoundedness given all lags, so the relevance of the test depends largely on the plausibility of the second assumption, stationarity and exchangeability.

6. Assessing, and Addressing Lack of, Overlap in Covariate Distributions

The second of the key assumptions in estimating average treatment effects requires that the propensity score is strictly between zero and one. Although in principle this is testable, as it restricts the joint distribution of observables, formal tests are not the main concern. In practice, this assumption raises a number of issues. The first question is how to detect a lack of overlap in the covariate distributions. A second is how to deal with it, given that such a lack exists.

6.1 Assessing Overlap in Covariate Score Distributions

The first method to assess overlap is to report some summary statistics for all covariates. Specifically, it is useful to report the normalized difference in covariate means by treatment status:

$$\text{nor}-\text{dif} = \frac{\overline{X}_1 - \overline{X}_0}{S^2_{X,0} + S^2_{X,1}},$$

where

$$\overline{X}_w = \frac{1}{N_w}\sum_{i:W_i=w} X_i \qquad \text{and} \quad S^2_{X,w} = \frac{1}{N_w - 1}\sum_{i:W_i=w}\left(X_i - \overline{X}_w\right)^2.$$

Note that we do not report the t-statistic for the difference,

$$t = \frac{\overline{X}_1 - \overline{X}_0}{S^2_{X,0}/N_0 + S^2_{X,1}/N_1}.$$

Essentially the t-statistic is equal to the normalized difference multiplied by the square root of the sample size. As such, the t-statistic partly reflects the sample size. Given a difference of 0.25 standard deviations between the two groups in terms of average covariate values, a larger t-statistic just indicates a larger sample size, and therefore in fact an easier problem in

terms of finding credible estimators for average treatment effects. As this example illustrates, a larger t-statistic for the difference between average covariates by treatment group does not indicate that the problem of finding credible estimates of the treatment effect is more difficult. A larger normalized difference does unambiguously indicate a more severe overlap problem.

In general a difference in average means bigger than 0.25 standard deviations is substantial. In that case one may want to be suspicious of simple methods like linear regression with a dummy for the treatment variable. Recall that estimating the average effect essentially amounts to using the controls to estimate the conditional mean $\mu_0(x) = \mathbb{E}[Y_i|W_i = 0, X_i = x]$ and using this estimated regression function to predict the (missing) control outcomes for the treated units. With such a large difference between the two groups in covariate distributions, linear regression is going to rely heavily on extrapolation, and thus will be sensitive to the exact functional form.

More generally one can plot distributions of covariates by treatment groups. In the case with one or two covariates one can do this directly. In high dimensional cases, however, this becomes more difficult. One can inspect pairs of marginal distributions by treatment status, but these are not necessarily informative about lack of overlap. It is possible that for each covariate the distribution for the treatment and control groups are identical, even though there are areas where the propensity score is zero or one.

A more direct method is to inspect the distribution of the propensity score in both treatment groups, which can reveal lack of overlap in the multivariate covariate distributions. Its implementation requires nonparametric estimation of the propensity score, however, and misspecification may lead to failure in detecting a lack of overlap, just as inspecting various marginal distributions may be insufficient. In practice one may wish to undersmooth the estimation of the propensity score, either by choosing a bandwidth smaller than optimal for nonparametric estimation or by including higher order terms in a series expansion.

6.2 Selecting a Sample with Overlap Through Matching

Once one determines that there is a lack of overlap one can attempt to construct a sample

with more overlap. Here we discuss two methods for doing so. The first is particularly appropriate when the focus is on the average effect for the treated, and there is a relatively large number of controls.

First, the treated observations are ordered, typically by decreasing values of the estimated propensity score. The reason for this is that among units with high values of the propensity score there are relatively more treated than control units, and therefore treated observations with high values of the propensity score are relatively more difficult to match.

Then the first treated unit (e.g., the one with the highest value for the estimated propensity score) is matched to the nearest control unit. Next, the second treated unit is matched to the nearest control unit, excluding the control unit that was used as a match for the first treated unit. Matching without replacement all treated units in this manner leads to a sample of $2 \cdot N_1$ units, (where N_1 is the size of the original treated subsample), half of them treated and half of them control units. Note that the matching is not necessarily used here as the final analysis. We do not propose to estimate the average treatment effect for the treated by averaging the differences within the pairs. Instead, this is intended as a preliminary analysis, with the goal being the construction of a sample with more overlap. Given a more balanced sample, one can use methods discussed in these notes for estimating the average effect of the treatment, including regression, propensity score methods, or matching. Using those methods on the balanced sample is likely to reduce bias relative to using the simple difference in averages by treatment status.

6.3 Selecting a Sample with Overlap Through Trimming

The second method for addressing lack of overlap we discuss is based on the work by Crump, Hotz, Imbens and Mitnik (2008). Their starting point is the definition of average treatment effects for subsets of the covariate space. Let $\mathbb{X}$ be the covariate space, and $\mathbb{A} \subset \mathbb{X}$ be some subset. Then define

$$\tau(\mathbb{A}) = \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\} \cdot \tau(X_i) \Big/ \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\}.$$

Crump et al calculate the efficiency bound for $\tau(\mathbb{A})$, assuming homoskedasticity, as

$$\frac{\sigma^2}{q(\mathbb{A})} \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1)}{1-e(X)} \middle| X \in \mathbb{A} \right],$$

where $q(\mathbb{A}) = \Pr(X \in \mathbb{A})$. They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance and show that it has the form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(X) \leq 1-\alpha\},$$

dropping observations with extreme values for the propensity score, with the cutoff value α determined by the equation

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X) \cdot (1-e(X))} \middle| \frac{1}{e(X) \cdot (1-e(X))} \leq \frac{1}{\alpha \cdot (1-\alpha)} \right].$$

Crump et al then suggest estimating $\tau(\mathbb{A}^*)$. Note that this subsample is selected solely on the basis of the joint distribution of the treatment indicators and the covariates, and therefore does not introduce biases associated with selection based on the outcomes. Calculations for Beta distributions for the propensity score suggest that $\alpha = 0.1$ approximates the optimal set well in practice.

7. Algorithm for Estimating the Propensity Score and Stratification

Many of the estimators discussed in this and the previous lecture are rely on estimators for the propensity score. Here I briefly describe one way of selecting a specification for the

propensity score. This particular procedure is a step-wise method, where increasingly flexible specifications are selected until the specification is deemed adequate. This is not the only way of doing, this, and in fact there are many such methods out there, some of which are undoubtedly more effective. The main point is that the common practice of including the full vector of covariates linearly, and not include any second order terms is not likely to be effective.

The algorithm starts with a K-dimensional vector of covariates X_i (these may already contain functions of the original covariates). The algorithm will selection a subset of the covariates to be included linearly, and based on that subset also select a number of second order terms (both quadratic terms and interactions).

The algorithm starts with a logistic model with no covariates. Next, logistic regression models are estimated with each of the covariates included separately. The covariate that improves the log likelihood function the most is included, as long as the increase in the log likelihood function is above some threshold t_{lin}. Next, we select among the $K-1$ remaining covariates the one that improves the logistic model with the single covariate the most, again based on the increase in the log likelihood function. We repeat this till no additional covariate improves the log likelihood function by at least t_{lin}. Suppose this leads to selecting $0 \leq K_{\text{lin}} \leq K$ covariates out of the original set of K covariates.

In the second part we select among the $K_{\text{lin}} \times (K_{\text{lin}}+1)/2$ second order terms based on these K_{lin} covariates. Similar to the way we selected the linear terms, we keep adding second order terms, till no additional second order term improves the log likelihood by more than t_{qua}. The tuning constants used below are $t_{\text{lin}} = 0.5$ and $t_{\text{qua}} = 1.35$, based on cutoffs for likelihood ratio test statistics (equal to twice the increaese in the log likelihood function) of 1 and 2.71, the latter corresponding to a 10% level test.

One may modify this algorithm by selecting a subset of the covariates to be included irrespective of the correlations with the treatment. In the analyses below, we selected the last pre-program earnings and the indicator for those earnings being positive to be included

in this way, prior to selecting further covariates.

Some of the estimators discussed below also require an algorithm for choosing the number and boundaries for the blocks. Here is the algorithm used below. We start with a single stratum. The option is to split the stratum in two equal parts, with the new boundary point the median of the values of the estimated propensity score in the old stratum. The old stratum will be split if three conditions are satisfied. First, the t-statistic for testing equality of the average estimated propensity score among treated and controls is at least 1.96, the number of treated and control observations in both new strata is at least 3, and the number of obsersvations in each block is at least 3 plus the dimension of the covariate vector X_i. We then keep splitting the strata in the middle, until none of the strata satisfies the criteria for further division.

8. An Illustration Based on the Lalonde Data

Here we look at application of the ideas discussed in these notes. We take the NSW job training data orinally collected by Lalonde (1986), and subsequently analyzed by Dehejia and Wahba (1999). These data are available on Dehejia's website (reference). The starting point is an experimental evaluation of this training program. Lalonde then constructed non-experimental comparison groups to investigate the ability of various econometric techniques to replicate the experimental results. In the current illustration we use three subsamples, the (experimental) trainees, the experimental controls, and a CPS comparison group. In both cases we focus on estimating the average effect of the treatment for the treated.

In the next three subsections we do the design part of the analysis. Without using the outcome data we first assess the overlap in covariate distributions, then assess whether strong ignorability has some credibility and finally create a matched sample and assess these issues there.

8.1 Summary Statistics

First we give some summary statistics

Table 1: Summary Statistics for Experimental Sample

	Trainees (N=260)		Controls (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	nor-dif	mean	(s.d.)	nor-dif
	260.00	0.00	185.00	0.00	0.00	0.00	0.00	0.00
Black	0.84	0.36	0.83	0.38	0.03	0.07	0.26	1.72
Hispanic	0.06	0.24	0.11	0.31	0.12	0.07	0.26	0.04
Age	25.82	7.16	25.05	7.06	0.08	33.23	11.05	0.56
Married	0.19	0.39	0.15	0.36	0.07	0.71	0.45	0.87
No Degree	0.71	0.46	0.83	0.37	0.21	0.30	0.46	0.64
Education	10.35	2.01	10.09	1.61	0.10	12.03	2.87	0.48
Earnings '74	2.10	4.89	2.11	5.69	0.00	14.02	9.57	1.11
Unempl '74	0.71	0.46	0.75	0.43	0.07	0.12	0.32	1.05
Earnings '75	1.53	3.22	1.27	3.10	0.06	13.65	9.27	1.23
Unempl. '75	0.60	0.49	0.68	0.47	0.13	0.11	0.31	0.84

In this table we report averages and standard deviations for the three subsamples. In addition we report for both the trainee/experimental-control and for the trainee/CPS-comparison-group pairs the normalized difference in average covariate values by treatment status, normalized by the standard deviation of these covariates:

$$\frac{\overline{X}_1 - \overline{X}_0}{\sqrt{S^2_{X,0} + S^2_{X,1}}}.$$

Again, it is not the statistical significance of this difference we are interested in, as much as the degree of difficulty of the statistical problem of adjusting for these differences. In Table 1 we see that in the experimental data set the difference in average age between treated and controls is 0.08 standard deviations. In the nonexperimental comparison the difference in age is 0.56 standard deviations.

Right away we can see that the experimental data set is well balanced. The difference in averages between treatment and control group is never more than 0.21 standard deviations. In contrast, with the CPS comparison group the differences between the averages are up

to 1.23 standard deviations from zero, suggesting there will be serious issues in obtaining credible estimates of the average effect of the treatment.

In Figures 1 and 2 we present histogram estimates of the distribution of the propensity score for the treatment and control group in the experimental Lalonde data. These distributions again suggest that there is considerable overlap in the covariate distributions. In Figures 3 and 4 we present the histogram estimates for the propensity score distributions for the CPS comparison group. Now there is a clear lack of overlap. For the CPS comparison group almost all mass of the propensity score distribution is concentrated in a small interval to the right of zero, and the distribution for the treatment group is much more spread out.

The results so far already strongly indicate that simple analyses such as least squares regression are unlikely to lead to credible estimates of the average causal effects of interest.

8.2 Assessing Unconfoundedness

First we use the experimental data. We analyze the data as if earnings in 1975 (Earn '75) is the pseudo outcome. This is in fact a covariate, and so it cannot be affected by the treatment, and we are looking for estimates that are substantially close to zero, and statistically indistinguishable from zero. Table 2 reports the results for nine estimators.

1. The first is the simple difference in average outcomes:

$$\hat{\tau} = \overline{Y}_1 - \overline{Y}_0.$$

2. The second estimator is based on least squares regression using all ten covariates:

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i.$$

3. The third estimator is based on least squares regression using all ten covariates and their interaction with the treatment indicator:

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \gamma'(X_i - \overline{X}_1) \cdot W_i + \varepsilon_i.$$

The interaction is based on deviations from the average covariate values for the treated in order for the least squares estimator for τ to estimate the average effect on the treated.

4. The fourth estimator uses the estimated propensity score to weight the observations:

$$\hat{\tau} = \frac{1}{N_1} \cdot \sum_{i:W_i=1} Y_i - \sum_{i:W_i=0} Y_i \cdot \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)} \Big/ \sum_{j:W_j=0} \frac{\hat{e}(X_j)}{1-\hat{e}(X_j)}.$$

The weights here are modified from those discussed previously to take account of the focus on the average effect for the treated.

5. Here the propensity score is used to create strata. Within the J strata the average effect is estimated as the difference in average outcomes between treated and controls, and the within-stratum estimates are averaged, weighted by the number of treated units in each strata. The number of strata is choosen in a data-dependent way, as described in Section 4.

6. Here all the treated observations are matched to the closets control, with replacement. The matching is on all covariates, weighted by the diagonal matrix with the inverse of the variances on the diagonal.

7. The seventh estimator is based on weighted least squares regression of the regression function

$$Y_i = \alpha + \tau \cdot W_i + \beta' X_i + \varepsilon_i,$$

with weights

$$\lambda_i = W_i + (1-W_i) \cdot \frac{\hat{e}(X_i)}{1-\hat{e}(X_i)}.$$

(The fourth estimator is a special case of this where β is set equal to zero.)

8. The eight estimator is based on the same blocks as the fifth estimator, but now within blocks linear regression is used to estimate the average effect.

TABLE 2: ESTIMATES FOR LALONDE DATA WITH EARNINGS '75 AS OUTCOME

	Experimental Controls			CPS Comparison Group		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	0.27	0.31	0.87	-12.12	0.25	-48.91
OLS (parallel)	0.22	0.22	1.02	-1.13	0.36	-3.17
OLS (separate)	0.17	0.22	0.74	-1.10	0.36	-3.07
Weighting	0.29	0.30	0.96	-1.56	0.26	-5.99
Blocking	0.26	0.32	0.83	-12.12	0.25	-48.91
Matching	0.11	0.25	0.44	-1.32	0.34	-3.87
Weighting and Regression	0.21	0.22	0.99	-1.58	0.23	-6.83
Blocking and Regression	0.12	0.21	0.59	-1.13	0.21	-5.42
Matching and Regression	-0.01	0.25	-0.02	-1.34	0.34	-3.96

9. The ninth estimator uses the same matching as the sixth estimator. Then linear regression is used on the 185 matches to estimate

$$Y_i = \alpha + \beta' X_i + \varepsilon_i,$$

and the estimated regression coefficients $\hat{\beta}$ are used to adjust the matched outcomes based on the Abadie-Imbens estimator.

For all nine estimators the estimated effect is close to zero and statistically insignificant at conventional levels. The results suggest that unconfoundedness is plausible for the experimental data set. This is not surprising, as the randomization implies unconfoundedness.

With the CPS comparison group the results are very different. All nine estimators suggest substantial and statistically significant differences in earnings in 1975 after adjusting for all other covariates, including earnings in 1974. This suggests that relying on the unconfoundedness assumption, in combination with these particular estimators, is not very credible for this sample. This is not surprising, because the treated and control samples

are so far apart, as measured by the normalized differences, that the estimates were very unlikely to be robust.

8.3 Creating a Matched Sample

Now let us consider the matched CPS sample. Matching is done on here the estimated propensity score, without replacement, for all the treated observations, starting with the treated unit with the highest value for the estimated propensity score. This leads to a matched sample with 185 treated (as before), and 185 controls. First we assess the balance by looking at the summary statistics.

Table 4: Summary Statistics for Matched CPS Sample

	Trainees (N=185)		Controls (N=185)		
	mean	(s.d.)	mean	(s.d.)	nor-dif
Black	0.84	0.36	0.85	0.35	-0.02
Hispanic	0.06	0.24	0.06	0.25	-0.02
Age	25.82	7.16	25.88	7.65	-0.01
Married	0.19	0.39	0.25	0.43	-0.10
No Degree	0.71	0.46	0.57	0.50	0.20
Education	10.35	2.01	10.91	2.93	-0.16
Earnings '74	2.10	4.89	2.81	5.61	-0.10
Unempl '74	0.71	0.46	0.66	0.47	0.07
Earnings '75	1.53	3.22	1.82	3.79	-0.06
Unempl. '75	0.60	0.49	0.50	0.50	0.14

These suggest that the balance is much improved, with the largest differences now on the order or 0.20 of a standard deviation, where before they difference was as high as 1.12. Now the normalized differences are comparable to those in the experimental sample.

Figures 5 and 6 present histograms of the propensity score for this matched sample. Note that we re-estimate the propensity score for this sample. If the matching had been perfect, the estimated propensity score would be equal to 0.5 for all units. It is not, and there is still considerable variation in the propensity score, but not to the extent that simple analyses

could not adjust for the covariate differences between the treatment and control samples.

These normalized differences suggest that given unconfoundedness, the matched sample is well balanced, and likely to lead to robust estimates. They do not directly reflect, however, on the question whether unconfoundedness itself is plausible. In order to address that, we return to the analysis with earnings in 1975 as the pseudo outcome. Again we report estimates for nine estimators. Here we do not directly use the matched sample from Table 4. Rather, we take the covariates excluding earnings in 1975 and the indicator for earnings in 1975 being positive, and create a matched sample based on the remaining covariates. This obviously makes the subsequent comparison more "fair". Based on this matched sample we re-estimate the effect of the treatment on the pseudo outcome, earnings in 1975. The results are in Table 5.

Table 5: Estimates on Matched CPS Lalonde Data

	Earn '75 Outcome			Earn '78 Outcome		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	-1.72	0.46	-3.74	0.87	0.80	1.08
OLS (parallel)	-1.51	0.33	-4.52	1.40	0.77	1.81
OLS (separate)	-1.40	0.32	-4.38	1.26	0.77	1.64
Weighting	-1.29	0.46	-2.80	1.20	0.80	1.49
Blocking	-1.30	0.47	-2.75	1.16	0.82	1.41
Matching	-1.50	0.39	-3.83	1.53	0.95	1.61
Weight and Regr	-1.38	0.33	-4.16	1.32	0.78	1.69
Block and Regr	-1.47	0.33	-4.41	1.77	0.76	2.33
Match and Regr	-1.51	0.39	-3.85	1.41	0.95	1.49

The results for earnings in 1975 still suggest substantial and statistically significant effects, so based on this we would **not** conclude that unconfoundedness is reasonable. Estimates are robust across the nine estimators.

Finally we report the estimates for earnings in 1978. Only now do we actually use the outcome data. Note that with the exclusion of the simple difference $\overline{Y}_1 - \overline{Y}_0$, the estimates

are all between 1.16 and 1,77, and thus relatively insensitive to the choice of estimator. The benchmark estimate from the experimental sample is $\overline{Y}_1 - \overline{Y}_0 = 1.79$, very similar to these non-experimental estimates. The irony is that all the estimators give answers consistent with the experimental estimates, but the analysis based on the pseudo outcome suggests we would not have known that without having the experimental estimates.

References

Blundell, R. and M. Costa-Dias (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics,"Institute for Fiscal Studies, Cemmap working paper cwp10/02.

Chen, X., H. Hong, and Tarozzi, (2005), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects,"unpublished working paper, Department of Economics, New York University.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2006), "Moving the Goalposts: Addressing Limited Overlap in Estimation of Average Treatment Effects by Changing the Estimand,"unpublished manuscript, Department of Economics, UC Berkeley.

Crump, R., V. J. Hotz, V. J., G. Imbens, and O.Mitnik, (2007), "Nonparametric Tests for Treatment Effect Heterogeneity,"forthcoming, *Review of Economics and Statistics*.

Dehejia, R. (2005) "Program Evaluation as a Decision Problem,"*Journal of Econometrics*, 125, 141-173.

Engle, R., D. Hendry, and J.-F. Richard, (1983) "Exogeneity,"*Econometrica*, 51(2): 277-304.

Firpo, S. (2003), "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259-276.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., H. Ichimura, and P. Todd, (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65, 261–294.

Heckman, J., R. Lalonde, and J. Smith (2000), "The Economics and Econometrics

of Active Labor Markets Programs,"in A. Ashenfelter and D. Card eds. Handbook of Labor Economics, vol. 3. New York: Elsevier Science.

Heckman, J., and J. Hotz, (1989), "Alternative Methods for Evaluating the Impact of Training Programs", (with discussion), *Journal of the American Statistical Association.*, Vol. 84, No. 804, 862-874.

Hirano, K., G. Imbens, and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189. July

Hirano, K., and J. Porter, (2005), "Asymptotics for Statistical Decision Rules,"Working Paper, Dept of Economics, University of Wisconsin.

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions,"*Biometrika*, Vol. 87, No. 3, 706-710.

Imbens, G., (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,"*Review of Economics and Statistics*, 86(1): 1-29.

Imbens, G., and J. Wooldridge., (2009), "Recent Developments in the Econometrics of Program Evaluation,"*Journal of Economic Literature*, Vol. Volume 47, Issue 1, 586.

Lalonde, R.J., (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data,"*American Economic Review*, 76, 604-620.

Lechner, M., (2001), "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption,"in Lechner and Pfeiffer (eds.), *Econometric Evaluations of Active Labor Market Policies in Europe*, Heidelberg, Physica.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects,"*American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., (2004), "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72(4), 1221-1246.

Manski, C. (2005), *Social Choice with Partial Knowledge of Treatment Response*, Princeton University Press.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Robins, J., and Y. Ritov, (1997), "Towards a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models," *Statistics in Medicine* 16, 285-319.

Rosenbaum, P., (1987), "The role of a second control group in an observational study", *Statistical Science*, (with discussion), Vol 2., No. 3, 292–316.

Rosenbaum, P., (1995), *Observational Studies*, Springer Verlag, New York.

Rosenbaum, P., and D. Rubin, (1983a), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", emphBiometrika, 70, 41-55.

Rosenbaum, P., and D. Rubin, (1983b), "Assessing the Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212-218.

Rubin, D., (1973a), "Matching to Remove Bias in Observational Studies", *Biometrics*, 29, 159-183.

Rubin, D., (1973b), "The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies", *Biometrics*, 29, 185-203.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D., (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal*

of Educational Statistics, 2(1), 1-26.

Rubin, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization", *A*nnals of Statistics, 6:34–58.

Rubin, D., (1990), "Formal Modes of Statistical Inference for Causal Effects", emphJournal of Statistical Planning and Inference, 25, 279-292.

Figure 1: histogram propensity score for controls, exper full sample

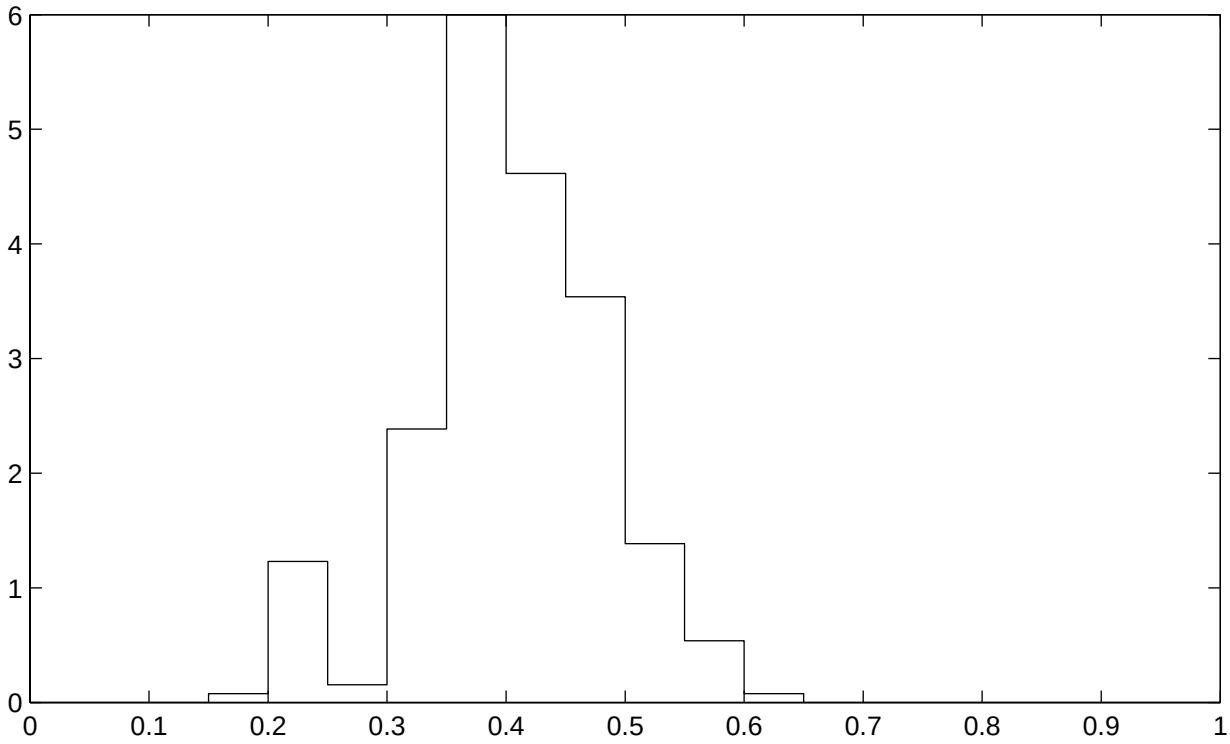


Figure 2: histogram propensity score for treated, exper full sample

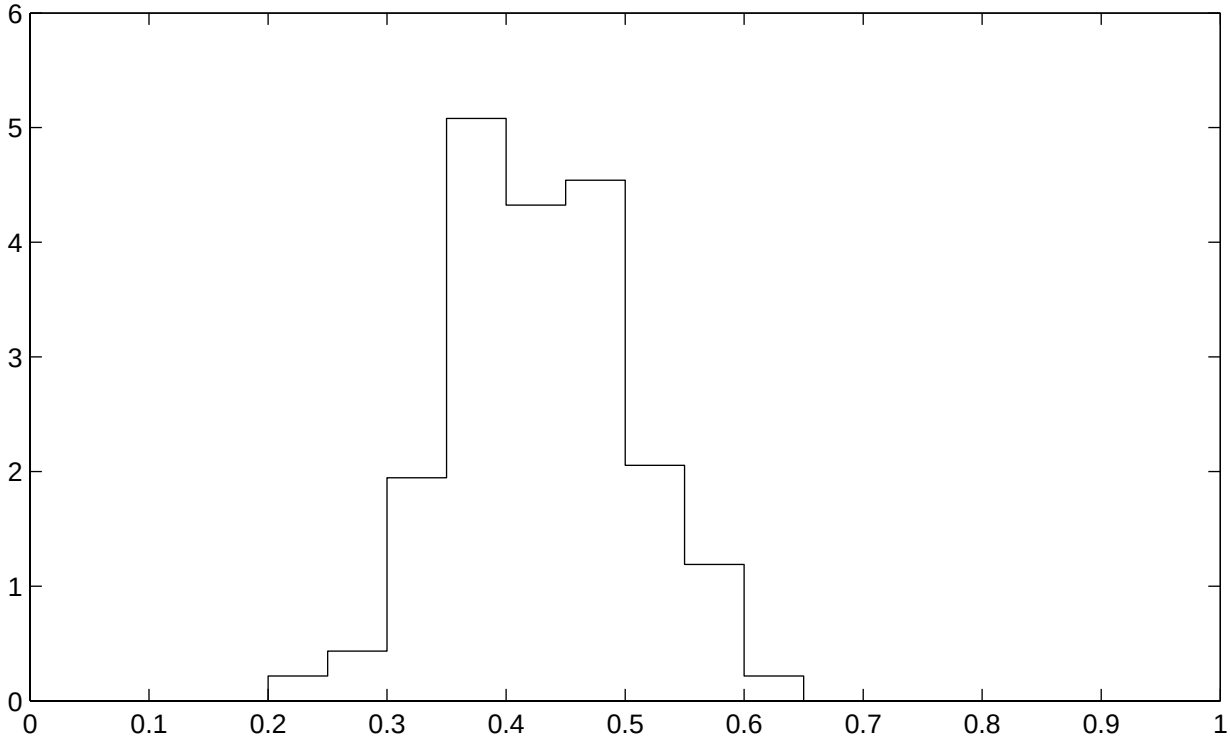


Figure 3: hist p−score for controls, cps full sample

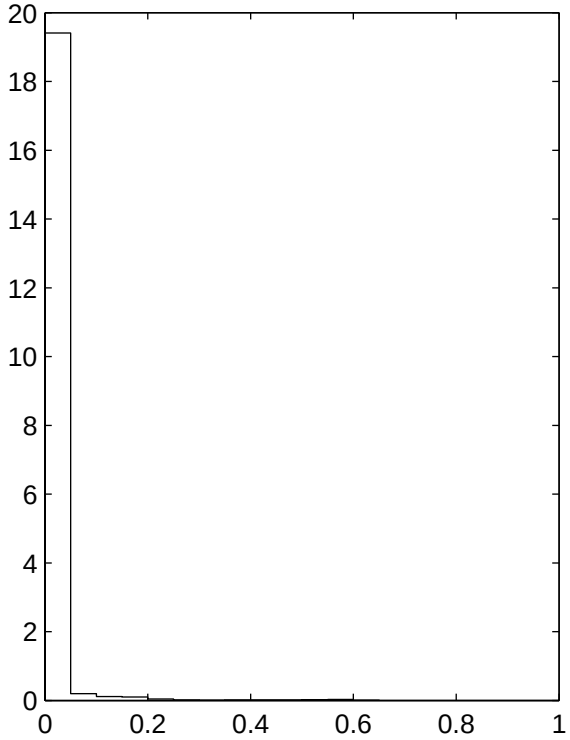


Figure 4: hist p−score for treated, cps full sample

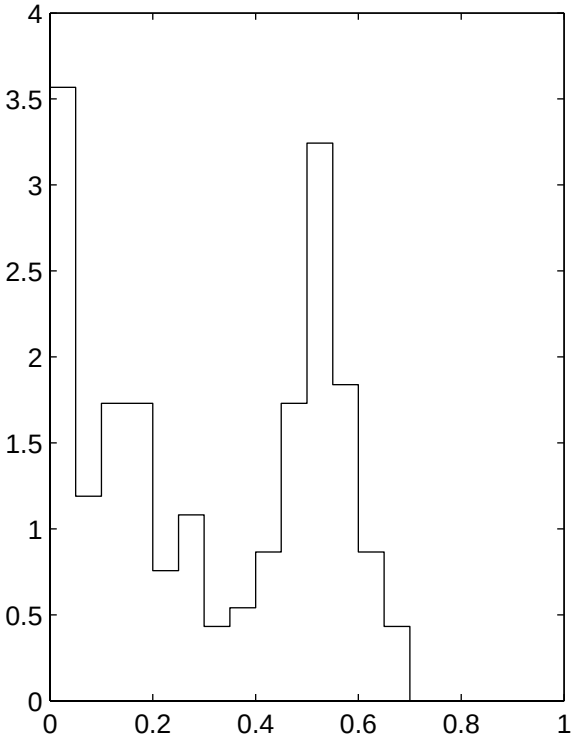


Figure 5: hist p−score for controls, cps selected sample

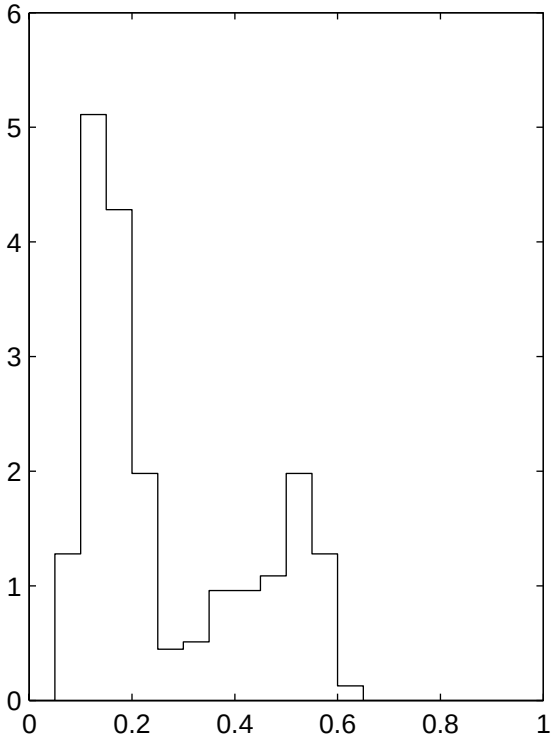
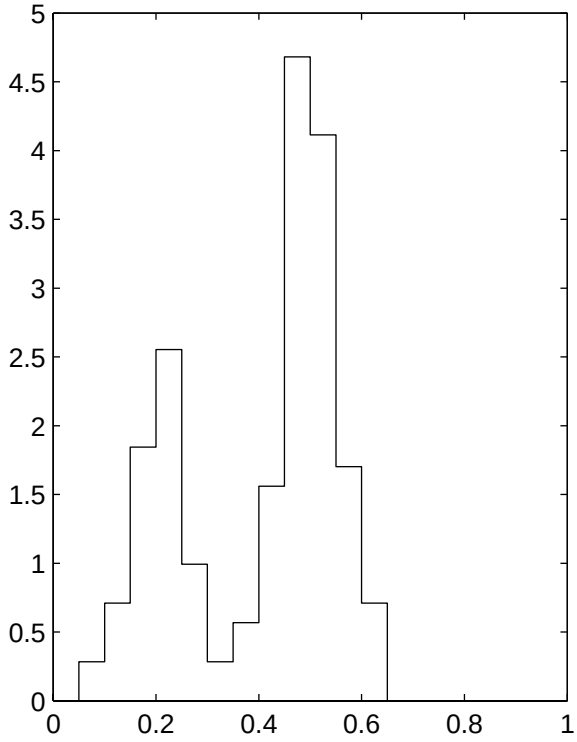


Figure 6: hist p−score for treated, cps selected sample



"Cross-Section Econometrics"

Lecture 1

Estimation of Average Treatment Effects

Under Unconfoundedness

Guido Imbens

AEA Lectures, Chicago, January 2012

Outline

1. Introduction
2. Potential Outcomes
3. Estimands and Identification
4. Estimation and Inference
5. Assessing Unconfoundedness (not testable)
6. Dealing with Overlap Problems
7. Illustration based on Lalonde Data

1. Introduction

We are interested in estimating the average effect of a program or treatment, allowing for heterogenous effects, assuming that selection can be taken care of by adjusting for differences in observed covariates.

This setting is of great applied interest.

Long literature, in both statistics and economics. Influential economics/econometrics papers include Ashenfelter and Card (1985), Barnow, Cain and Goldberger (1980), Card and Sullivan (1988), Dehejia and Wahba (1999), Hahn (1998), Heckman and Hotz (1989), Heckman and Robb (1985), Lalonde (1986). In stat literature work by Rubin (1974, 1978), Rosenbaum and Rubin (1983).

Unusual case with many proposed (semi-parametric) estimators (matching, regression, propensity score, or combinations), many of which are actually used in practice.

We discuss implementation, and assessment of the critical assumptions (even if they are not testable).

In practice concern with overlap in covariate distributions tends to be important.

Once overlap issues are addressed, choice of estimators is less important. Estimators combining matching and regression or weighting and regression are recommended for robustness reasons.

Key role for analysis of the joint distribution of treatment indicator and covariates prior to using outcome data.

2. Potential Outcomes (Rubin, 1974)

We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population.

We postulate the existence for each unit of a pair of potential outcomes,
$Y_i(0)$ for the outcome under the control treatment and
$Y_i(1)$ for the outcome under the active treatment
$Y_i(1) - Y_i(0)$ is unit-level causal effect

Covariates X_i (not affected by treatment)
Each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment and $W_i = 1$ if unit i receives the active treatment. We observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Several additional pieces of notation.

First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

$$e(x) = \mathrm{Pr}(W_i = 1 | X_i = x) = \mathbb{E}[W_i | X_i = x].$$

Also the two conditional regression and variance functions:

$$\mu_w(x) = \mathbb{E}[Y_i(w) | X_i = x], \quad \sigma_w^2(x) = \mathbb{V}(Y_i(w) | X_i = x).$$

3 Estimands, Assumptions, and Identification

3.A Estimands

Population average treatments

$$\tau_P = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \tau_{P,T} = \mathbb{E}[Y_i(1) - Y_i(0) | W = 1].$$

Most of the discussion in these notes will focus on τ_P, with extensions to $\tau_{P,T}$ available in the references.

We will also look at the sample average treatment effect (SATE):

$$\tau_S = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

τ_P versus τ_S does not matter for estimation, but matters for variance.

3.B. Assumptions

Assumption 1 (Unconfoundedness, Rosenbaum and Rubin, 1983a)

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i.$$

"conditional independence assumption," "selection on observables." In missing data literature "missing at random."

To see the link with standard exogeneity assumptions, assume constant effect and linear regression:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i, \quad \Longrightarrow \quad Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. Given the constant treatment effect assumption, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous.

Motivation for Unconfoundeness Assumption (I)

The first is a statistical, data descriptive motivation.

A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units.

A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment).

Such an analysis may not lead to the final word on the efficacy of the treatment, but the absence of such an analysis would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

Motivation for Unconfoundeness Assumption (II)

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not.

The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated.

It is clear that settings where some of necessary covariates are not observed will require strong assumptions to allow for identification. E.g., instrumental variables settings Absent those assumptions, typically only bounds can be identified (e.g., Manski, 1990, 1995).

Motivation for Unconfoundeness Assumption (III)

Example of a model that is consistent with unconfoundedness: suppose we are interested in estimating the average effect of a binary input on a firm's output, or $Y_i = g(W, \varepsilon_i)$.

Suppose that profits are output minus costs,

$$W_i = \arg\max_w \mathbb{E}[\pi_i(w)|c_i] = \arg\max_w \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon_i) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of output given inputs, then unconfoundedness will hold as

$$(g(0, \varepsilon_i), g(1, \varepsilon_i)) \perp\!\!\!\perp c_i.$$

Overlap

Second assumption on the joint distribution of treatments and covariates:

Assumption 2 (Overlap)

$$0 < \Pr(W_i = 1|X_i) < 1.$$

Rosenbaum and Rubin (1983a) refer to the combination of the two assumptions as "stongly ignorable treatment assignment."

3.C Identification

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$$

$$= \mathbb{E}[Y_i(1)|X_i = x, W_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, W_i = 0]$$

$$= \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0].$$

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y_i|X_i = x, W_i = w]$ for all values of w and x in the support of these variables. This is where overlap is important.

Given identification of $\tau(x)$,

$$\tau_P = \mathbb{E}[\tau(X_i)]$$

Alternative Assumptions

$$\mathbb{E}[Y_i(w)|W_i, X_i] = \mathbb{E}[Y_i(w)|X_i],$$

for $w = 0, 1$. Although this assumption is unquestionably weaker, in practice it is rare that a convincing case can be made for the weaker assumption without the case being equally strong for the stronger Assumption.

The reason is that the weaker assumption is intrinsically tied to functional form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (e.g., logarithms) without the strong assumption.

If we are interested in $\tau_{P,T}$ it is sufficient to assume

$$Y_i(0) \perp\!\!\!\perp W_i \mid X_i,$$

3.D The Role of the Propensity Score

Result 1 *Suppose that Assumption 1 holds. Then:*

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid e(X_i).$$

Only need to condition on scalar function of covariates, which would be much easier in practice if X_i is high-dimensional.

(Problem is that the propensity score $e(x)$ is almost never known.)

4. Estimation and Inference

4.A Efficiency Bound

Hahn (1998): for any regular estimator for τ_P, denoted by $\widehat{\tau}$, with

$$\sqrt{N} \cdot (\widehat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}(0, V),$$

the variance must satisfy:

$$V \geq \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\tau(X_i) - \tau_P)^2\right]. \quad (1)$$

Estimators exist that achieve this bound.

4.B Estimators

A. Regression Estimators

B. Matching

C. Propensity Score Estimators

D. Mixed Estimators **(recommended)**

4.B.1 Regression Estimators

Estimate $\mu_w(x)$ consistently and estimate τ_P or τ_S as

$$\widehat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)).$$

Simple implementations include

$$\mu_w(x) = \beta' x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ. In this case one can estimate τ simply by least squares estimation using the regression function

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i.$$

More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$.

These simple regression estimators can be sensitive to differences in the covariate distributions for treated and control units.

The reason is that in that case the regression estimators rely heavily on extrapolation.

Note that $\mu_0(x)$ is used to predict missing outcomes for the treated. Hence on average one wishes to use predict the control outcome at $\overline{X}_T = \sum_i W_i \cdot X_i / N_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \widehat{\beta}'(\overline{X}_T - \overline{X}_C)$.

If $\overline{X}_T$ and $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. With the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.

4.B.2 Matching

let $\ell_m(i)$ is the mth closest match, that is, the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\} = m,$$

Then

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if } W_i = 1, \end{cases} \qquad \widehat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j & \text{if} \\ Y_i & \text{if} \end{cases}$$

The simple matching estimator is

$$\widehat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right). \qquad (2)$$

Issues with Matching

Bias is of order $O(N^{-1/K})$, where K is dimension of covariates. Is important in large samples if $K \geq 2$ (and dominates variance asymptotically if $K \geq 3$)

Not Efficient (but efficiency loss is small)

Easy to implement, robust.

4.B.3 Propensity Score Estimators: Weighting

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY_i(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e(X)Y_i(1)}{e(X)}\right]\right] = \mathbb{E}[Y_i(1)],$$

and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y_i(0)],$$

implying

$$\tau_P = \mathbb{E}\left[\frac{W \cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i \cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right). \tag{3}$$

Implementation of Horvitz-Thompson Estimator

Estimate $e(x)$ flexibly (Hirano, Imbens and Ridder, 2003)

$$\widehat{\tau}_{\text{weight}} = \sum_{i=1}^{N}\frac{W_i \cdot Y_i}{\widehat{e}(X_i)} \Big/ \sum_{i=1}^{N}\frac{W_i}{\widehat{e}(X_i)}$$

$$- \sum_{i=1}^{N}\frac{(1-W_i)\cdot Y_i}{1-\widehat{e}(X_i)} \Big/ \sum_{i=1}^{N}\frac{(1-W_i)}{1-\widehat{e}(X_i)}$$

Is efficient given nonparametric estimator for $e(x)$.

Potentially sensitive to estimator for propensity score.

4.B.4 Blocking on the Propensity Score

Divide the sample into M blocks of units of approximately equal probability of treatment. Letting J_{im} be an indicator for unit i being in block m:

$$J_{im} = 1\{(m-1)/M < e(X_i) \leq m/M\},$$

and let $N_{wm} = \sum_i 1\{W_i = w, J_{im} = 1\}$. Estimate within each block the average treatment effect as if random assignment holds,

$$\widehat{\tau}_m = \frac{1}{N_{1m}}\sum_{i=1}^{N} J_{im}W_iY_i - \frac{1}{N_{0m}}\sum_{i=1}^{N} J_{im}(1-W_i)Y_i.$$

Then estimate the overall average treatment effect as:

$$\widehat{\tau}_{\text{block}} = \sum_{m=1}^{M}\widehat{\tau}_m \cdot \frac{N_{1m}+N_{0m}}{N}.$$

Matching or Regression on the Estimated Propensity Score

Used widely in practice.

Large sample properties not known: probably variance estimates based on ignoring estimation error in propensity score are conservative in large samples (as in weighting estimator case).

4.B.5 Mixed Estimators: Weighting and Regression

Interpret Horvitz-Thompson estimator as weighted regression estimator:

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i, \quad \text{with weights } \lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}}.$$

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example as

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i. Such an estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are specified correctly. That is, in the Robins-Ritov terminology, the estimator is doubly robust.
still relies on global approximations

4.B.6 Matching and Regression

First match observations.

Define

$$\widehat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)} & \text{if } W_i = 1, \end{cases} \qquad \widehat{X}_i(1) = \begin{cases} X_{\ell(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

Then adjust within pair difference for the within-pair difference in covariates $\widehat{X}_i(1) - \widehat{X}_i(0)$:

$$\widehat{\tau}_M^{\text{adj}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\beta} \cdot \left(\widehat{X}_i(1) - \widehat{X}_i(0) \right) \right),$$

using regression estimate for β.

Can eliminate bias of matching estimator given flexible specification of regression function.

4.B.7 Blocking and Regression

Rosenbaum and Rubin (1983b) suggest modifying the basic blocking estimator by using least squares regression within the blocks. Without the additional regression adjustment the estimated treatment effect within blocks can be written as a least squares estimator of τ_m for the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \varepsilon_i,$$

using only the units in block m. As above, one can also add covariates to the regression function

$$Y_i = \alpha_m + \tau_m \cdot W_i + \beta_m' X_i + \varepsilon_i,$$

again estimated on the units in block m.

4.C Estimation of the Variance

For efficient estimator of τ_P:

$$V_P = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \tau)^2\right],$$

Estimate all components nonparametrically, and plug in.

Alternatively, use bootstrap.

(Does not work for matching estimator)

Estimation of the Variance

For all estimators of τ_S, for some known $\lambda_i(\mathbf{X}, \mathbf{W})$

$$\widehat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

$$V\left(\widehat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma^2_{W_i}(X_i).$$

To estimate $\sigma^2_{W_i}(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $v(i)$ be the closest unit to i with the same treatment indicator.

The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

$$\widehat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{v(i)}\right)^2/2.$$

5.A Assessing Unconfoundedness: Multiple Control Groups

Suppose we have a three-valued indicator $T_i \in \{-0, 1, 1\}$ for the groups (e.g., ineligibles, eligible nonnonparticipants and participants), with the treatment indicator equal to $W_i = 1\{T_i = 1\}$, so that

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i \in \{-1, 0\} \\ Y_i(1) & \text{if } T_i = 1. \end{cases}$$

Suppose we extend the unconfoundedness assumption to independence of the potential outcomes and the three-valued group indicator given covariates,

$$Y_i(0), Y_i(1) \perp\!\!\!\perp T_i \,\Big|\, X_i$$

Now a testable implication is

$$Y_i(0) \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\},$$

and thus

$$Y_i \perp\!\!\!\perp 1\{T_i = 0\} \,\Big|\, X_i, T_i \in \{-1, 0\}.$$

An implication of this independence condition is being tested by the tests discussed above. Whether this test has much bearing on the unconfoundedness assumption, depends on whether the extension of the assumption is plausible given unconfoundedness itself.

5.B Assessing Unconfoundedness: Estimate Effects on Pseudo Outcomes

Partition the covariate vector into $X_i = (X_i^p, X_i^r)$, X_i^p scalar.

Unconfoundedness assumes

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid (X_i^p, X_i^r)$$

Suppose we are willing to assume X_i^r is sufficient:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i^r$$

and suppose X_i^p is a good proxy for $Y_i(0)$, then we can test

$$X_i^p \perp\!\!\!\perp W_i \mid X_i^r$$

Most useful implementations with X_i^p a lagged outcome.

Suppose the covariates consist of a number of lagged outcomes $Y_{i,-1}, \ldots, Y_{i,-T}$ as well as time-invariant individual characteristics Z_i, so that $X_i = (X_i^p, X_i^r)$, with $X_i^p = Y_{i,-1}$ and $X_i^r = (Y_{i,-2}, \ldots, Y_{i,-T}, Z_i)$. Outcome is $Y_i = Y_{i,0}$.

Now consider the following two assumptions. The first is unconfoundedness given only $T-1$ lags of the outcome:

$$Y_{i,0}(1), Y_{i,0}(0) \perp\!\!\!\perp W_i \mid Y_{i,-1}, \ldots, Y_{i,-(T-1)}, Z_i,$$

Then, under stationarity it seems reasonable to expect Then it follows that

$$Y_{i,-1} \perp\!\!\!\perp W_i \mid Y_{i,-2}, \ldots, Y_{i,-T}, Z_i,$$

which is testable.

6.A Assessing Overlap

The first method to detect lack of overlap is to look at summary statistics for the covariates by treatment group.

Most important here is the normalized difference in covariates:

$$\text{nor} - \text{dif} = \frac{\overline{X}_1 - \overline{X}_0}{S_{X,0}^2 + S_{X,1}^2}$$

$$\overline{X}_w = \frac{1}{N_w} \sum_{i:W_i=w} X_i \quad \text{and} \quad S_{X,w}^2 = \frac{1}{N_w - 1} \sum_{i:W_i=w} \left(X_i - \overline{X}_w\right)^2$$

Note that we do not report the t-statistic for the difference,

$$t = \frac{\overline{X}_1 - \overline{X}_0}{S_{X,0}^2/N_0 + S_{X,1}^2/N_1}$$

The t-statistic partly reflects the sample size. Given the normalized difference, a larger t-statistic just indicates a larger sample size, and therefore in fact an easier problem in terms of finding credible estimators for average treatment effects.

In general a difference in average means bigger than 0.25 standard deviations is substantial. In that case one may want to be suspicious of simple methods like linear regression with a dummy for the treatment variable.

Recall that estimating the average effect essentially amounts to using the controls to estimate $\mu_0(x) = \mathbb{E}[Y_i | W_i = 0, X_i = x]$ and using this estimated regression function to predict the (missing) control outcomes for the treated units.

With a large difference between the two groups, linear regression is going to rely heavily on extrapolation, and thus will be sensitive to the exact functional form.

6.B Assessing Overlap by Inspecting the Propensity Score Distribution

The second method for assessing overlap is more directly focused on the overlap assumption.

It involves inspecting the marginal distribution of the propensity score in both treatment groups.

Any difference in covariate distribution shows up in differences in the average propensity score between the two groups.

Moreoever, any area of non-overlap shows up in zero or one values for the propensity score.

6.C Selecting a Subsample with Overlap: Matching

Appropriate when the focus is on the average effect for treated, $\mathbb{E}[Y_i(1)-Y_i(0)|W_i=1]$, and when there is a relatively large pool of potential controls.

Order treated units by estimated propensity score, highest first.

Match highest propensity score treated unit to closest control on estimated propensity score, without replacement.

Only to create balanced sample, not as final analysis.

6.D Selecting a Subsample with Overlap: Trimming

Define average effects for subsamples $\mathbb{A}$:

$$\tau(\mathbb{A}) = \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\} \cdot \tau(X_i) / \sum_{i=1}^{N} 1\{X_i \in \mathbb{A}\}.$$

The efficiency bound for $\tau(\mathbb{A})$, assuming homoskedasticity, as

$$\frac{\sigma^2}{q(\mathbb{A})} \cdot \mathbb{E}\left[\frac{1}{e(X)} + \frac{1}{1-e(X)} \middle| X \in \mathbb{A}\right],$$

where $q(\mathbb{A}) = \Pr(X \in \mathbb{A})$.

They derive the characterization for the set $\mathbb{A}$ that minimizes the asymptotic variance .

The optimal set has the form

$$\mathbb{A}^* = \{x \in \mathbb{X} | \alpha \leq e(X) \leq 1-\alpha\},$$

dropping observations with extreme values for the propensity score, with the cutoff value α determined by the equation

$$\frac{1}{\alpha \cdot (1-\alpha)} = 2 \cdot \mathbb{E}\left[\frac{1}{e(X)\cdot(1-e(X))} \middle| \frac{1}{e(X)\cdot(1-e(X))} \leq \frac{1}{\alpha\cdot(1-\alpha)}\right].$$

Note that this subsample is selected solely on the basis of the joint distribution of the treatment indicators and the covariates, and therefore does not introduce biases associated with selection based on the outcomes.

Calculations for Beta distributions for the propensity score suggest that $\alpha = 0.1$ approximates the optimal set well in practice.

7. Applic. to Lalonde Data (Dehejia-Wahba Sample)

Data on job training program, first used by Lalonde (1986), See also Heckman and Hotz (1989), Dehejia and Wahba (1999).

Small experimental evaluation, 185 trainees, 260 controls, group of very disadvantaged in labor market.

Large, non-experimental comparison group from CPS (15,992 observations). Very different in distribution of covariates.

How well do the non-experimental results replicate the experimental ones? Is non-experimental analysis credible? Would we have known whether it was credible without experiments results?

Table 1: Summary Statistics for Lalonde Data

	Trainees (N=260)		Controls (N=185)			CPS (N=15,992)		
	mean	(s.d.)	mean	(s.d.)	n-dif	mean	(s.d.)	n-dif
Black	0.84	0.36	0.83	0.38	0.03	0.07	0.26	1.72
Hispanic	0.06	0.24	0.11	0.31	0.12	0.07	0.26	0.04
Age	25.8	7.2	25.1	7.1	0.08	33.2	11.1	0.56
Married	0.19	0.39	0.15	0.36	0.07	0.71	0.45	0.87
No Deg	0.71	0.46	0.83	0.37	0.21	0.30	0.46	0.64
Educ	10.4	2.0	10.1	1.6	0.10	12.0	2.9	0.48
Earn '74	2.10	4.89	2.11	5.69	0.00	14.02	9.57	1.11
U '74	0.71	0.46	0.75	0.43	0.07	0.12	0.32	1.05
Earn '75	1.53	3.22	1.27	3.10	0.06	13.65	9.27	1.23
U '75	0.60	0.49	0.68	0.47	0.13	0.11	0.31	0.84

The experimental data set is well balanced. The difference in averages between treatment and control group is never more than 0.21 standard deviations.

In contrast, with the CPS comparison group the differences between the averages are up to 1.23 standard deviations from zero, suggesting there will be serious issues in obtaining credible estimates of the average effect of the treatment.

Next, let us assess unconfoundedness in this sample using earnings in 1975 as the pseudo outcome.

We report results for 9 different estimators, including the simple difference, parallel and separate least squares regressions, weighting and blocking on the propensity score, and matching, with the last three also combined with regression.

Both for experimental control group and for cps comparison group.

Specification for propensity score, and block choice are based on algorithm, see notes for details.

Table 2: Estimates for Lalonde Data with Earnings '75 as Outcome

	Experimental Controls			CPS Comparison Group		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	0.27	0.31	0.87	-12.12	0.25	-48.91
OLS (parallel)	0.22	0.22	1.02	-1.13	0.36	-3.17
OLS (separate)	0.17	0.22	0.74	-1.10	0.36	-3.07
Weighting	0.29	0.30	0.96	-1.56	0.26	-5.99
Blocking	0.26	0.32	0.83	-12.12	0.25	-48.91
Matching	0.11	0.25	0.44	-1.32	0.34	-3.87
Weight and Regr	0.21	0.22	0.99	-1.58	0.23	-6.83
Block and Regr	0.12	0.21	0.59	-1.13	0.21	-5.42
Match and Regr	-0.01	0.25	-0.02	-1.34	0.34	-3.96

With the cps comparison group, results are discouraging. Consistently find big "effects" on earnings in 1975, with point estimates varying widely.

The sensitivity is not surprising given substantial differences in covariate distributions.

Next, create a matched sample to improve balance.

Order treated observations on estimated propensity score.

Starting with the highest propensity score, match each treated observation to the closest control, without replacement. Match on the propensity score.

Table 4: Summary Statistics for Matched CPS Sample

	Trainees (N=185)		Controls (N=185)		
	mean	(s.d.)	mean	(s.d.)	nor-dif
Black	0.84	0.36	0.85	0.35	-0.02
Hispanic	0.06	0.24	0.06	0.25	-0.02
Age	25.82	7.16	25.88	7.65	-0.01
Married	0.19	0.39	0.25	0.43	-0.10
No Degree	0.71	0.46	0.57	0.50	0.20
Education	10.35	2.01	10.91	2.93	-0.16
Earnings '74	2.10	4.89	2.81	5.61	-0.10
Unempl '74	0.71	0.46	0.66	0.47	0.07
Earnings '75	1.53	3.22	1.82	3.79	-0.06
Unempl. '75	0.60	0.49	0.50	0.50	0.14

In the matched sample the normalized differences are comparable to those in the experimental sample.

Now we revisit the matching analysis using earnings in 1975 as the pseudo outcome, and also carry out analysis on earnings in 1978 (the actual outcome).

Table 5: Estimates on Matched CPS Lalonde Data

	Earn '75 Outcome			Earn '78 Outcome		
	est	(s.e.)	t-stat	est	(s.e.)	t-stat
Simple Dif	-1.72	0.46	-3.74	0.87	0.80	1.08
OLS (parallel)	-1.51	0.33	-4.52	1.40	0.77	1.81
OLS (separate)	-1.40	0.32	-4.38	1.26	0.77	1.64
Weighting	-1.29	0.46	-2.80	1.20	0.80	1.49
Blocking	-1.30	0.47	-2.75	1.16	0.82	1.41
Matching	-1.50	0.39	-3.83	1.53	0.95	1.61
Weight and Regr	-1.38	0.33	-4.16	1.32	0.78	1.69
Block and Regr	-1.47	0.33	-4.41	1.77	0.76	2.33
Match and Regr	-1.51	0.39	-3.85	1.41	0.95	1.49

The results for earnings in 1975 still suggest substantial and statistically significant effects, so based on this we would **not** conclude that unconfoundedness is reasonable. Estimates are robust across the nine estimators.

Estimates for earnings in 1978 are robust accross all nine estimators, with the exception of the simple difference in average outcomes by treatment status. These Estimates are consistent with experimental estimates (1.77).

Conclusion

Important to assess and address lack of overlap.

In reasonably balanced samples choice of estimator is less important.

Combining regression and matching or propensity score blocking is preferred method for robustness properties.

AEA Lectures **Chicago, IL, January 2012**

Lecture 3, Sunday, Jan 7th,pm-pm

Selection on Unobservables:

Part I: Instrumental Variables, Local Average Treatment Effects

1. Introduction

In this lecture we discuss the interpretation of instrumental variables estimators allowing for general heterogeneity in the effect of the endogenous regressor. We shall see that instrumental variables estimators generally estimate average treatment effects, with the specific average depending on the choice of instruments. Initially we focus on the case where both the instrument and the endogenous regressor are binary. The example we will use is based on one of the best known examples of instrumental variables, the paper by Joshua Angrist on estimating the effect of veteran status on earnings (Angrist, 1990). We also discuss the case where the instrument and or the endogenous variable take on multiple values, and incorporate the presence of covariates.

The general theme of this lecture is that with heterogenous treatment effects, endogeneity creates severe problems for identification of population averages. Population average causal effects are only estimable under very strong assumptions on the effect of the instrument on the endogenous regressor (sometimes referred to as "identification at infinity", Chamberlain, 1986), or under the constant treatment effect assumptions. Without such assumptions we can only identify average effects for subpopulations that are induced by the instrument to change the value of the endogenous regressors. Following Angrist, Imbens and Rubin (1996), we refer to such subpopulations as *compliers*, and we refer to the average treatment effect that is point identifed as the *local average treatment effect* (Imbens and Angrist, 1994). The "complier" terminology stems from the canonical example of a randomized experiment with noncompliance. In this example a random subpopulation is assigned to the treatment, but some of the individuals do not comply with their assigned treatment.

These complier subpopulations are not necessarily the subpopulations that are *ex ante* the

most interesting subpopulations. The reason to nevertheless focus on these subpopulations is that the data are generally not informative about average effects for other subpopulations without extrapolation, similar to the way in which a randomized experiment conducted on men is not informative about average effects for women without extrapolation. The set up here allows the researcher to sharply separate the extrapolation to the (sub-)population of interest, from exploration of the information in the data about the causal effect of interest. The latter analysis relies primarily on relatively interpretable, and substantively meaningful assumptions, and it avoids functional form or distributional assumptions. Subsequently, given estimates for the compliers, one can these estimates in combination with the data to assess the plausibility of extrapolating the local average treatment effect to other subpopulations, using the information on outcomes given one of the two treatment levels and covariates, or construct bounds on the average effects for the primary population of interest using the bounds approach from Manski (e.g., Manski, 2008).

With multiple instruments, and/or with covariates, one can assess the evidence for heterogeneity, and therefore investigate the plausibility of extrapolation to the full population more extensively.

2. Linear Instrumental Variables with Constant Coefficients

First let us briefly review standard textbook linear instrumental variables methods (e.g., Wooldridge, 2000). In the example from Angrist (1990) we use to illustrate the concepts discussed in this lecture we are interested in the causal effect of military service on earnings, using eligibility for the draft as the instrument. Let Y_i be the outcome of interest for unit i (log earnings in the example), W_i the binary endogenous regressor (an indicator for veteran status), and Z_i the binary instrument (a binary indicator for draft eligibility). The standard set up is as follows. A linear model is postulated for the relation between the outcome and the endogenous regressor:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \varepsilon_i. \tag{1}$$

This is a structural, behavioral, or causal relationship (we use the terms interchangeably).

The concern is that the regressor W_i is endogenous, that is, that W_i is correlated with the unobserved component of the outcome, ε_i. Suppose that we are confident that a second observed covariate, the instrument Z_i, is both uncorrelated with the unobserved component ε_i and correlated with the endogenous regressor W_i. The solution is to use Z_i as an instrument for W_i. There are a couple of ways to implement this.

In Two-Stage-Least-Squares (TSLS) we first estimate a linear regression of the endogenous regressor on the instrument by least squares. Let the estimated regression function be

$$\hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 \cdot Z_i.$$

Then we regress the outcome on the predicted value of the endogenousr regressor, using least squares:

$$\hat{Y}_i = \hat{\alpha} + \hat{\tau}^{\text{tsls}} \cdot \hat{W}_i.$$

Alternatively, with a single instrument we can estimate the two reduced form regressions

$$\hat{Y}_i = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot Z_i, \qquad \text{and} \quad \hat{W}_i = \hat{\pi}_0 + \hat{\pi}_1 \cdot Z_i,$$

by least squares and estimate β_1 through Indirect Least Squares (ILS) as the ratio

$$\hat{\tau}^{\text{ils}} = \hat{\gamma}_1 / \hat{\pi}_1,$$

irrespective of the validity of the behavioral model.

In the case with a single instrument and single endogenous regressor, we end up in both cases with the ratio of the sample covariance of Y_i and Z_i to the sample covariance of W_i and Z_i.

$$\hat{\tau}^{\text{iv}} = \hat{\tau}^{\text{ils}} = \hat{\tau}^{\text{tsls}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}) \cdot (Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^{N}(W_i - \bar{W}) \cdot (Z_i - \bar{Z})}.$$

This estimator is consistent for

$$\tau^{\text{iv}} = \frac{\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]}{\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]}. \tag{2}$$

Using a central limit theorem for all the moments and the delta method we can infer the large sample distribution without additional assumptions:

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{iv}} - \tau^{\text{iv}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2 \cdot (Z_i - \mathbb{E}[Z_i])^2\right]}{\left(\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]\right)^2}\right),$$

where $\varepsilon_i = Y_i - \mathbb{E}[Y_i] - \tau^{\text{iv}} \cdot (W_i - \mathbb{E}[W_i])$. Under independence between the residual ε_i and the instrument Z_i, the asymptotic distribution further simplifies to:

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{iv}} - \tau^{\text{iv}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[\varepsilon_i^2\right] \cdot \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])^2\right]}{\left(\mathbb{E}\left[(W_i - \mathbb{E}[W_i]) \cdot (Z_i - \mathbb{E}[Z_i])\right]\right)^2}\right),$$

3. Potential Outcomes

First we set up the problem in a slightly different way, using Rubin's (1974) potential outcomes approach to causality. This set up of the instrumental variables problem originates with Imbens and Angrist (1994). Let $Y_i(0)$ and $Y_i(1)$ be two potential outcomes for unit i, one for each value of the endogenous regressor or treatment. The first potential outcome $Y_i(0)$ measures the outcome if person i were not to serve in the military, irrespective of whether this person served or not. The second potential outcome, $Y_i(1)$, measures the outcome given military service, again irrespective of whether the person served or not. We are interested in the causal effect of military service, $Y_i(1) - Y_i(0)$. We cannot directly observe this since we can only observe either $Y_i(0)$ or $Y_i(1)$, never both. Let W_i be the realized value of the endogenous regressor, equal to zero or one. We observe W_i and

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 1 \\ Y_i(0) & \text{if } W_i = 0. \end{cases}$$

So far the set up is identical to that in the analysis under unconfoundedness in Lecture 1. Now we introduce additional notation by defining similar potential outcomes for the treatment.

Initially we focus on the case with a binary instrument Z_i. In the Angrist example, Z_i is a binary indicator for having a draft number below the cutoff value that implied a potential recruit would get called up for military service, and thus an indicator for being draft eligible. Define two potential outcomes $W_i(0)$ and $W_i(1)$, representing the value of the endogenous regressor given the two values for the instrument. The actual or realized (and observed) value of the endogenous variable is

$$W_i = Y_i(Z_i) = \begin{cases} W_i(1) & \text{if } Z_i = 1 \\ W_i(0) & \text{if } Z_i = 0. \end{cases}$$

In summary, we observe the triple (Z_i, W_i, Y_i), where $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i(Z_i))$.

4. Local Average Treatment Effects

In this section we interpret the estimand (2) under weaker assumptions than the linear additive model set up in (1).

4.1. Assumptions

The key instrumental variables assumption is

Assumption 1 (Independence)

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), W_i(0), W_i(1)).$$

This assumption requires that the instrument is as good as randomly assigned, and that it does not directly affect the outcome. The assumption is formulated in a nonparametric way, without definitions of residuals that are tied to functional forms.

It is important to note that this assumption is *not* implied by random assignment of Z_i. To see this, an alternative formulation of the assumption, slightly generalizing the notation, is useful. First we postulate the existence of four potential outcomes, $Y_i(z, w)$, corresponding to the outcome that would be observed if the instrument was exogenously set to $Z_i = z$ and the treatment was exogenously set to $W_i = w$. Then the independence assumption is the combination of two assumptions.

Assumption 2 (Random Assignment)

$$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1)).$$

Assumption 3 (Exclusion Restriction)

$$Y_i(z,w) = Y_i(z',w), \qquad \text{for all } z, z', w.$$

The first of these two assumptions is implied by random assignment of Z_i. It can be weakened in the presence of covariates to unconfoundedness. The second assumption is substantive, and randomization has no bearing on it. It corresponds to the notion that there is no direct effect of the instrument on the outcome other than through the treatment. In the model-based version of this, (1), it is captured by the absence of Z_i in the behavioral equation. This assumption has to be argued on a case-by-case basis.

It is useful for our approach to think about the compliance behavior of the different individuals or units, that is how they respond in terms of the treatment received to different values of the instrument. Table 1 gives the four possible pairs of values $(W_i(0), W_i(1))$, given the binary nature of the treatment and instrument and their labels. The labels refer to the canonical example of a randomized experiment with imperfect compliance.

Table 1: Compliance Types

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

We cannot directly establish the type of an individual based on what we observe for them (the triple Z_i, W_i, Y_i)) since we only see the pair (Z_i, W_i), not the pair $(W_i(0), W_i(1))$

(typically observing Y_i is immaterial for this argument). Nevertheless, we can rule out some possibilities. Table 2 summarizes the information about compliance behavior from observed treatment status and instrument. For each pair of (Z_i, W_i) values there are two possible

Table 2: Compliance Type by Treatment and Instrument

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

types, with the two others ruled out.

To make additional progress we we consider a *monotonicity* assumption, also known as the *no-defiers* assumption, introduced by Imbens and Angrist (1994):

Assumption 4 (Monotonicity/No-Defiers)

$$W_i(1) \geq W_i(0).$$

This monotonicity assumption is very apealling in many applications. It is implied directly by many (constant coefficient) latent index models of the type:

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \varepsilon_i > 0\}, \tag{3}$$

which would imply $W_i(1) \geq W_i(0)$ if $\pi_1 \geq 0$ and $W_i(1) \leq W_i(0)$ otherwise. In the canonical example of a randomized experiment with non-compliance this assumption is very plausible: if Z_i is assignment to a treatment, and W_i is an indicator for receipt of treatment, it makes sense that there are few, if any, individuals who always to the exact opposite of what their assignment is.

4.2. The Local Average Treatment Effect

Given monotonicity we can infer more about an individual's compliance behavior, as summarized in Table 3. For individuals with (Z_i, W_i) equal to $(0, 1)$ or $(1, 0)$ we can now

Table 3: Compliance Type by Treatment and Instrument given Monotonicity

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

determine their type. For individuals with (Z_i, W_i) equal to $(0, 0)$ or $(1, 1)$ there are still multiple types consistent with the observed behavior. Nevertheless, we can stochastically infer the compliance types.

Now we proceed to identifying the marginal distribution of types and conditional potential outcome distributions. Let π_c, π_n, and π_a be the population proportions of compliers, never-takers and always-takers respectively. We can identify those from the population distribution of treatment and instrument status:

$$\mathbb{E}[W_i|Z_i = 0] = \pi_a, \qquad \mathbb{E}[W_i|Z_i = 1] = \pi_a + \pi_c,$$

which we can invert to infer the population shares of the different types:

$$\pi_a = \mathbb{E}[W_i|Z_i = 0], \qquad \pi_c = \mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0],$$

and

$$\pi_n = 1 - \pi_a - \pi_c = 1 - \mathbb{E}[W_i|Z_i = 1].$$

Now consider average outcomes by instrument and treatment status. In the (Z_i, W_i) equal to $(0,1)$ or $(1,0)$ subpopulations these expectations have a simple interpretation:

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 1] = \mathbb{E}[Y_i(0)|\text{never} - \text{taker}], \tag{4}$$

and

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 0] = \mathbb{E}[Y_i(1)|\text{always} - \text{taker}]. \tag{5}$$

In the (Z_i, W_i) equal to $(0,0)$ or $(1,1)$ the conditional outcome expectations are mixtures of expected values for compliers and nevertakers and compliers and alwaystakers respectively:

$$\mathbb{E}[Y_i|W_i = 0, Z_i = 0] = \frac{\pi_c}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{complier}] + \frac{\pi_n}{\pi_c + \pi_n} \cdot \mathbb{E}[Y_i(0)|\text{never} - \text{taker}], \tag{6}$$

and

$$\mathbb{E}[Y_i|W_i = 1, Z_i = 1] = \frac{\pi_c}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{complier}] + \frac{\pi_a}{\pi_c + \pi_a} \cdot \mathbb{E}[Y_i(1)|\text{always} - \text{taker}]. \tag{7}$$

From these relationships we can infer the average outcome by treatment status for compliers, first by combining (4) and (6),

$$\mathbb{E}[Y_i(0)|\text{complier}] = \frac{\pi_c + \pi_n}{\pi_n} \cdot \mathbb{E}[Y_i|W_i = 0, Z_i = 0] - \frac{\pi_c}{\pi_n} \cdot \mathbb{E}[Y_i|W_i = 0, Z_i = 1],$$

and then by combining (5) and (7)

$$\mathbb{E}[Y_i(1)|\text{complier}] = \frac{\pi_c + \pi_a}{\pi_a} \cdot \mathbb{E}[Y_i|W_i = 1, Z_i = 1] - \frac{\pi_c}{\pi_a} \cdot \mathbb{E}[Y_i|W_i = 1, Z_i = 0].$$

Thus we can infer the average effect for compliers, $\mathbb{E}[Y(1) - Y_i(0)|\text{complier}] = \mathbb{E}[Y_i(1)|\text{complier}] - \mathbb{E}[Y_i(0)|\text{complier}]$.

It turns out this is equal to the instrumental variables estimand (2). Consider the least squares regression of Y_i on a constant and Z_i. The slope coefficient in that regression estimates

$$\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0].$$

The two terms are equal to:

$$\mathbb{E}[Y_i|Z_i=1] = \mathbb{E}[Y_i(1)|\text{complier}]\cdot\pi_c + \mathbb{E}[Y_i(0)|\text{never}-\text{taker}]\cdot\pi_0 + \mathbb{E}[Y_i(1)|\text{always}-\text{taker}]\cdot\pi_a.$$

and

$$\mathbb{E}[Y_i|Z_i=0] = \mathbb{E}[Y_i(0)|\text{complier}]\cdot\pi_c + \mathbb{E}[Y_i(0)|\text{never}-\text{taker}]\cdot\pi_0 + \mathbb{E}[Y_i(1)|\text{always}-\text{taker}]\cdot\pi_a.$$

Hence the difference is

$$\mathbb{E}[Y_i|Z_i=1] - \mathbb{E}[Y_i|Z_i=0] = \mathbb{E}[Y_i(1)-Y_i(0)|\text{complier}]\cdot\pi_c.$$

The same argument can be used to show that the slope coefficient in the regression of W_i on Z_i is

$$\mathbb{E}[W_i|Z_i=1] - \mathbb{E}[W_i|Z_i=0] = \pi_c.$$

Hence the instrumental variables estimand, the ratio of these two reduced form estimands, is equal to the local average treatment effect

$$\beta^{\text{iv}} = \frac{\mathbb{E}[Y_i|Z_i=1] - \mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[W_i|Z_i=1] - \mathbb{E}[W_i|Z_i=0]} = \mathbb{E}[Y_i(1)-Y_i(0)|\text{complier}]. \tag{8}$$

The key insight is that the data are informative only about the average effect for compliers only. Put differently, the data are not informative about the average effect for nevertakers because nevertakers are never seen receiving the treatment, and they are not informative about the average effect for alwaystakers because alwaystakers are never seen without the treatment. A similar insight in a parametric settings is discussed in Björklund and Moffitt (1987). (These results do not take away from the fact that one can construct informative bounds about the average effect for nevertakers or alwaystakers based on the outcomes we do observe for such individuals, in the spirit of the work by Manski, 2008.)

A special case of considerable interest is that with one-side non-compliance. Suppose that $W_i(0) = 0$, so that those assigned to the control group cannot receive the active treatment (but those assigned to the active treatment can choose to receive it or not, so that $W_i(1) \in \{0, 1\}$). In that case only two compliance types remain, compliers and always-takers. Monotonicity is automatically satisfied, and the average effect for compliers is now equal to the average effect for the treated, since any one receiving the treatment is by definition a complier. This case was first studied in Bloom (1984). It also has a useful connection to Chamberlain's notion of "identification at infinity," (see also Heckman, 1990). Suppose that we have a selection model with a participation equation as in (3), with $\pi_1 > 0$. If Z_i is a continuous instrument, then in order to idenfify the average effect for the treated we need Z_i to have unbounded support. Within this specific selection model this is, as Chamberlain (1987) in a different context, an unattractive identification condition. However, in many application it is plausible that there is some value of the instrument such that individuals do not have access to the treatment, implying identification of the average effect for the treated.

4.3 Extrapolating to the Full Population

Although we cannot consistently estimate the average effect of the treatment for always-takers and never-takers, we do have some information about the potential outcomes for these subpopulations that can aid in assessing the plausibility of extrapolating to average effects for the full population. They key insight is that we can infer the average outcome for never-takers and always-takers in one of the two treatment arms. Specifically, we can estimate

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \tag{9}$$

but not

$$\mathbb{E}\left[Y_i(1)|\text{never}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(0)|\text{always}-\text{taker}\right],$$

We can learn from the expectations in (9) whether there is any evidence of heterogeneity in

outcomes by compliance status, by comparing the pair of average outcomes of $Y_i(0)$;

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(0)|\text{complier}\right],$$

and the pair of average outcomes of $Y_i(1)$:

$$\mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \qquad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{complier}\right].$$

If compliers, never-takers and always-takers are found to be substantially different in levels, based on evidence of substantial difference between $\mathbb{E}[Y_i(0)|\text{never}-\text{taker}]$ and $\mathbb{E}[Y_i(0)|\text{complier}]$, and or/between $\mathbb{E}[Y_i(1)|\text{always}-\text{taker}]$, and $\mathbb{E}[Y_i(1)|\text{complier}]$, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types. On the other hand, if one finds that outcomes given the control treatment for never-takers and compliers are similar, and outcomes given the treatment are similar for compliers and always-takers (and especially if this holds within various subpopulations defined by observed covariates), then it appears to be more plausible that average treatment effects for these groups are also comparable.

4.4 Covariates

The local average treatment effect result in (8) implies in general that one cannot consistently estimate average effects for subpopulations other than compliers. This still holds in cases where we observe covariates. One can incorporate the covariates into the analysis in a number of different ways. Traditionally the TSLS or ILS set up is used with the covariates entering in the structural outcome equation linearly and additively, as

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \beta_2' X_i + \varepsilon_i,$$

with the covariates added to the set of instruments. Given the potential outcome set up with general heterogeneity in the effects of the treatment, one may also wish to allow for more heterogeneity in the correlations with the covariates. Here we describe a general way of doing so. Unlike TSLS-type approaches, this involves modelling both the dependence of

the outcome and the treatment on the covariates. Although there is often a reluctance to model the relation between the treatment, there is no apparent reason that economic theory is more informative about the relation between covariates and outcomes than about the relation between covariates and the choices that lead to the treatment.

A full model can be decomposed into two parts, a model for the compliance type given covariates, and a model for the potential outcomes given covariates for each compliance type. A traditional parametric model with a dummy endogenous variables might have the form (translated to the potential outcome set up used here):

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \pi_2' X_i + \eta_i \geq 0\}, \tag{10}$$

$$Y_i(w) = \beta_0 + \beta_1 \cdot w + \beta_2' X_i + \varepsilon_i, \tag{11}$$

with (η_i, ε_i) jointly normally distributed and independent of the instruments(e.g., Heckman, 1978). A more general model would allow for separate outcome equations by treatment status:

$$Y_i(0) = \beta_{00} + \beta_{20}' X_i + \varepsilon_{0i}, \tag{12}$$

$$Y_i(1) = \beta_{01} + \beta_{21}' X_i + \varepsilon_{1i}, \tag{13}$$

in combination with (10), (e.g., Björklund and Moffitt, 1987). Such models can be viewed as imposing various restrictions on the relation between compliance types, covariates and outcomes. For example, in the model characterized by equations (10) and (11), if $\pi_1 > 0$, compliance type depends on η_i:

$$\text{unit } i \text{ is a } \begin{cases} \text{never} - \text{taker} & \text{if } \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{complier} & \text{if } -\pi_0 - \pi_1 - \pi_2' X_i \leq \eta_i < -\pi_0 - \pi_1 - \pi_2' X_i \\ \text{always} - \text{taker} & \text{if } -\pi_0 - \pi_2' X_i \leq \eta_i. \end{cases}$$

Not only does this impose monotonicity, by ruling out the presence of defiers, it also implies strong restrictions on the relationship between type and outcomes. Specifically, the selection

equation implies that compliers correspond to intermediate values of η_i, implying that conditional expectations of $Y_i(0)$ and $Y_i(1)$ for compliers are in between those for never-takers and always-takers.

An alternative approach to the conventional selection model that exploits the identification results more directly, is to model the potential outcome $Y_i(w)$ for units with compliance type t given covariates X_i through a common functional form with type and treatment specific parameters:

$$f_{Y(w)|X,T}(y(w)|x,t) = f(y|x;\theta_{wt}),$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$. For example, using a normal model,

$$Y_i(w)|T_i = t, X_i = x \sim \mathcal{N}\left(x'\beta_{wt}, \sigma^2_{wt}\right), \tag{14}$$

for $(w,t) = (0,n),(0,c),(1,c),(1,a)$.

A natural model for the distribution of type is a trinomial logit model:

$$\text{pr}(T_i = \text{complier}|X_i) = \frac{1}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)},$$

$$\text{pr}(T_i = \text{never}-\text{taker}|X_i) = \frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)},$$

and

$$\text{pr}(T_i = \text{always}-\text{taker}|X_i) = 1 - \Pr(T_i = \text{complier}|X_i) - \Pr(T_i = \text{never}-\text{taker}|X_i).$$

The log likelihood function is then, factored in terms of the contribution by observed (W_i, Z_i) values, using the normal model for the conditional outcomes in (14):

$$\mathcal{L}(\pi_n, \pi_a, \beta_{0n}, \beta_{0c}, \beta_{1c}, \beta_{1a}, \sigma_{0n}, \sigma_{0c}, \sigma_{1c}, \sigma_{1a}) =$$

$$\times \prod_{i|W_i=0,Z_i=1} \frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{0n}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0n}}{\sigma_{0n}}\right)$$

$$\times \prod_{i|W_i=0,Z_i=0} \left(\frac{\exp(\pi_n' X_i)}{1+\exp(\pi_n' X_i)} \cdot \frac{1}{\sigma_{0n}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0n}}{\sigma_{0n}}\right) + \frac{1}{1+\exp(\pi_n' X_i)} \cdot \frac{1}{\sigma_{0c}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{0c}}{\sigma_{cn}}\right)\right)$$

$$\times \prod_{i|W_i=1,Z_i=1} \left(\frac{\exp(\pi_a' X_i)}{1+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1a}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1a}}{\sigma_{1a}}\right) + \frac{1}{1+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1c}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1c}}{\sigma_{1c}}\right)\right)$$

$$\times \prod_{i|W_i=1,Z_i=0} \frac{\exp(\pi_a' X_i)}{1+\exp(\pi_n' X_i)+\exp(\pi_a' X_i)} \cdot \frac{1}{\sigma_{1a}} \cdot \phi\left(\frac{Y_i - X_i'\beta_{1a}}{\sigma_{1a}}\right).$$

For example, the second factor consists of the contributions of individuals with $Z_i = 0$, $W_i = 0$, who are known to be either compliers or never-takers. Maximizing a likelihood function with this mixture structure is straightforward using the EM algorithm (Dempster, Laird, and Rubin, 1977). For an empirical example of this approach see Hirano, Imbens, Rubin and Zhou (2000), and Imbens and Rubin (1997).

In small samples one may wish to incorporate restrictions on the effects of the covariates, and for example assume that the effect of covariates on the potential outcome is the same irrespective of compliance type, or even irrespective of the treatment status. An advantage of this approach is that it can easily be generalized. The type probabilities are nonparametricaly identified as functions of the covariates, and the similarly the outcome distributions are nonparametrically identified, by type as a function of the covariates,.

5. Effects of Military Service on Earnings

In a classic application of instrumental variables methods Angrist (1989) was interested in estimating the effect of serving in the military on earnings. He was concerned about the possibility that those choosing to serve in the military are different from those who do not in ways that affects their subsequent earnings irrespective of serving in the military. To avoid biases in simple comparisons of veterans and non-veterans, he exploited the Vietnam era draft lottery. Specifically he uses the binary indicator whether or not someone's draft

lottery number made him eligible to be drafted as an instrument. The lottery number was tied to an individual's day of birth, so more or less random. Even so, that in itself does not make it valid as an instrument as we shall discuss below. As the outcome of interest Angrist uses total earnings for a particular year.

The simple ols regression leads to:

$$\begin{array}{rcccc} \widehat{\log(\text{earnings})}_i & = & 5.4364 & - & 0.0205 \cdot \widehat{\text{veteran}}_i \\ & & (0079) & & (0.0167) \end{array}$$

In Table 4 we present population sizes of the four treatment/instrument subsamples. For example, with a low lottery number 5,948 individuals do not, and 1,372 individuals do serve in the military.

Table 4: Treatment Status by Assignment

		Z_i	
		0	1
W_i	0	5,948	1,915
	1	1,372	865

Using these data we get the following proportions of the various compliance types, given in Table 5, under the no-defiers or monotonicity assumption. For example, the proportion of nevertakers is estimated as the conditional probability of $W_i = 0$ given $Z_i = 1$:

$$\text{pr(nevertaker)} = \frac{1915}{1915 + 865} = 0.6888.$$

Table 6 gives the average outcomes for the four groups, by treatment and instrument status.

Table 5: Compliance Types: Estimated Proportions

		$W_i(0)$	
		0	1
$W_i(1)$	0	never-taker (0.6888)	defier (0)
	1	complier (0.1237)	always-taker (0.1874)

Table 6: Estimated Average Outcomes by Treatment and Instrument

		Z_i	
		0	1
W_i	0	$\widehat{\mathbb{E}[Y]} = 5.4472$	$\widehat{\mathbb{E}[Y]} = 5.4028$
	1	$\widehat{\mathbb{E}[Y]} = 5.4076,$	$\widehat{\mathbb{E}[Y]} = 5.4289$

Table 7 gives the estimated averages for the four compliance types, under the exclusion restriction. This restriction is the key assumption here. There are a number of reasons why it may be violated in this application. For example, never-takers may need to taking active action to avoid military service if draft eligible, for example by continuing their formal education, or by moving to Canada. Always-takers may be affected their lottery number if draftees were treated differently in the military compared to volunteers. The local average treatment effect is -0.2336, a 23% drop in earnings as a result of serving in the military.

Simply doing IV or TSLS would give you the same numerical results:

$$\begin{array}{rcccc} \widehat{\log(\text{earnings})}_i & = & 5.4836 & - & 0.2336 \cdot \widehat{\text{veteran}}_i \\ & & (0.0289) & & (0.1266) \end{array}$$

It is interesting in this application to inspect the average outcome for different compli-

Table 7: Compliance Types: Estimated Average Outcomes

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker: $\widehat{\mathbb{E}[Y_i(0)]} = 5.4028$	defier (NA)
	1	complier: $\widehat{\mathbb{E}[Y_i(0)]} = 5.6948$, $\widehat{\mathbb{E}[Y_i(1)]} = 5.4612$	always-taker: $\widehat{\mathbb{E}[Y_i(1)]} = 5.4076$

ance groups. Average log earnings for never-takers are 5.40, lower by 29% than average earnings for compliers who do not serve in the military. This suggests that never-takers are substantially different than compliers, and that the average effect of 23% for compliers need not be informative never-takers. In contrast, average log earnings for always-takers are only 6% lower than those for compliers who serve, suggesting that the differences between always-takers and compliers are considerably smaller. Note that compliers have better outcomes without the treatment than never-takers and better outcomes than always-takers given the treatment. This is inconsistent with the simple normal selection model in(10)-(11).

6. Multivalued Instruments

For any two values of the instrument z_0 and z_1 satisfying the local average treatment effect assumptions we can define the corresponding local average treatment effect:

$$\tau_{z_1,z_0} = \mathbb{E}[Y_i(1) - Y_i(0)|W_i(z_1) = 1, W_i(z_0) = 0].$$

Note that these local average treatment effects need not be the same for different pairs of instrument values. Comparisons of estimates based on different instruments underlies tests of overidentifying restrictions in TSLS settings. An alternative interpretation of rejections in such testing procedures is therefore the presence of heterogeneity in causal effects, rather than that some of the instruments are invalid. Without restrictions on the heterogeneity of the causal effects there are no tests in general for the validity of the instruments.

The presence of multi-valued, or similarly, multiple, instruments, does, however, provide an opportunity to assess variation in treatment effects, as well as an opportunity to obtain average effects for subpopulations closer to the one of ultimate interest. Suppose that we have an instrument Z_i with support $z_0, z_1, \ldots, z_K$. Suppose also that the monotonicity assumption holds for all pairs z and z', and suppose that the instruments are ordered in such a way that

$$p(z_{k-1}) \leq p(z_k), \qquad \text{where} \quad p(z) = \mathbb{E}[W_i | Z_i = z].$$

Also suppose that the instrument is relevant, so that for some function $g(Z)$,

$$\mathbb{E}[g(Z_i) \cdot (W_i - \mathbb{E}[W_i])] \neq 0.$$

Then the instrumental variables estimator based on using $g(Z)$ as an instrument for W estimates a weighted average of the local average treatment effects $\tau_{z_k, z_{k-1}}$:

$$\tau_g = \frac{\text{Cov}(Y_i, g(Z_i))}{\text{Cov}(W_i, g(Z_i))} = \sum_{k=1}^{K} \lambda_k \cdot \tau_{z_k, z_{k-1}},$$

where the weights λ_k are non-negative and satisfy

$$\lambda_k = \frac{(p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]}{\sum_{k=1}^{K} p(z_k) - p(z_{k-1})) \cdot \sum_{l=k}^{K} \pi_l (g(z_l) - \mathbb{E}[g(Z_i)]},$$

for

$$\pi_k = \text{pr}(Z_i = z_k),$$

implying that $\sum_{k=1}^{K} \lambda_k = 1$.

Choosing the function $g(z)$ corresponds to choosing the weight function. There are obviously limits to the weight functions that can be choosen. One can only estimate a weighted average of the local average treatment effects defined for all pairs of instrument values in the support of the instrument. If $p(z_0) = 0$ for some z_0 in the support of Z, one can estimate the average effect on the treated as τ_{z_K, z_0}.

If the instrument Z has a continuous distribution, and the probability of receiving the treatment given the instrument, $p(z)$, is continuous in z, we can define the limit of the local average treatment effects

$$\tau_z = \lim_{z' \downarrow z, z'' \uparrow z} \tau_{z',z''}.$$

If the monotonicity assumption holds for all pairs z and z', we can use the implied structure on the compliance behavior by modelling $W_i(z)$ as a threshold crossing process,

$$W_i(z) = 1\{h(z) + \eta_i \geq 0\}, \tag{15}$$

with the scalar unobserved component η_i independent of the instrument Z_i. This type of latent index model is used extensively in work by Heckman (Heckman and Robb, 1985; Heckman,1990; Heckman and Vytlacil, 2005), as well as in Vytlacil (2000). Vytlacil shows that if the earlier three assumptions (independence, the exclusion restriction and monotonicity) hold for all pairs z and z', than there is a function $h(\cdot)$ such that this latent index structure is consistent with the joint distribution of the observables. The latent index structure implies that individuals can be ranked in terms of an unobserved component η_i such that if for two individuals i and j we have $\eta_i > \eta_j$, than $W_i(z) \geq W_j(z)$ for all z.

Given this assumption, we can define the marginal treatment effect $\tau(\eta)$ as

$$\tau(\eta) = \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| \eta_i = \eta\right].$$

In a parametric setting this was introduced by Björklund and Moffitt (1987). In the continuous Z case this marginal treatment effect relates directly to the limit of the local average treatment effects:

$$\tau(\eta) = \tau_z, \qquad \text{with} \quad \eta = -h(z)).$$

Note that we can only define $\tau(\eta)$ for values of η for which there is a z such that $\tau = -h(z)$. Normalizing the marginal distribution of η to be uniform on $[0, 1]$ (Vytlacil, 2002), this

restricts η to be in the interval $[\inf_z p(z), \sup_z p(z)]$, where $p(z) = \text{pr}(W_i = 1|Z_i = z)$. Heckman and Vytlacil (2005) characterize various average treatment effects (e.g., the population average treatment effect, the average treatment effect for the treated, the local average treatment effect) in terms of this marginal treatment effect. For example, the population average treatment effect is simply the average of the marginal treatment effect over the marginal distribution of η:

$$\tau = \int_\eta \tau(\eta) dF_\eta(\eta).$$

In practice the same limits remain on the identification of average effects. A necessary condition for identification of the population average effect is that the instrument moves the probability of participation from zero to one. Note that identification of the population average treatment effect does not require identification of $\tau(\eta)$ at every value of η. The latter is sufficient, but not necessary. For example, in a randomized experiment (corresponding to a binary instrument with the treatment indicator equal to the instrument) the population average treatment effect is obviously identified, but the marginal treatment effect is not identified for any value of η.

7. Multivalued Endogenous Variables

Now suppose that the endogenous variable W_i takes on values $0, 1, \ldots, J$. We still assume that the instrument Z_i is binary. We study the interpretation of the instrumental variables estimand

$$\tau^{\text{iv}} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(W_i, Z_i)} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[W_i|Z_i = 1] - \mathbb{E}[W_i|Z_i = 0]}.$$

We make the exclusion assumption that for all z in the support of Z_i,

$$Y_i(w), W_i(z) \perp\!\!\!\perp Z_i,$$

and a version of the monotonicity assumption,

$$W_i(1) \geq W_i(0).$$

Then we can write the instrumental variables estimand as

$$\tau^{\text{iv}} = \sum_{j=1}^{J} \lambda_j \cdot \mathbb{E}[Y_i(j) - Y_i(j-1) | W_i(1) \geq j > W_i(0)], \tag{16}$$

where

$$\lambda_j = \frac{\text{pr}(W_i(1) \geq j > W_i(0)}{\sum_{i=1}^{J} \text{pr}(W_i(1) \geq i > W_i(0)}. \tag{17}$$

The weights are non-negative and add up to one.

Note that we can estimate the weights λ_j because

$$\begin{aligned}
\text{pr}(W_i(1) \geq j > W_i(0) &= \text{pr}(W_i(1) \geq j) - \text{pr}(W_i(0) \geq j) \\
&= \text{pr}(W_i(1) \geq j | Z_i = 1) - \text{pr}(W_i(0) \geq j | Z_i = 0) \\
&= \text{pr}(W_i \geq j | Z_i = 1) - \text{pr}(W_i \geq j | Z_i = 0),
\end{aligned}$$

using the monotonicity assumption.

8. Instrumental Variables Estimates of the Returns to Education Using Quarter of Birth as an Instrument

Here we use a subset of the data used by Angrist and Krueger in their 1991 study of the returns to education. Angrist and Krueger were concerned with the endogeneity of education, with the standard argument that individuals with higher ability are likely to command higher wages at any level of education, as well as be more likely to choose high levels of education. In that case simple least squares estimates would over estimate the returns to education. Angrist and Krueger realized that individuals born in different parts of the year are subject to slightly different compulsory schooling laws. If you are born before a fixed cutoff date you enter school at a younger age than if you are born after that cutoff date, and given that you are allowed to leave school when you turn sixteen, those individuals born before the

cutoff date are required to completely more years of schooling. The instrument can therefore be thought of as the tightness of the compulsory schooling laws, with the tightness being measured by the individual's quarter of birth.

Angrist and Krueger implement this using census data with quarter of birth indicators as the instrument. Table 8 gives average years of education and sample sizes by quarter of birth.

Table 8: Average Level of Education by Quarter of Birth

quarter	1	2	3	4
average level of education	12.69	12.74	12.81	12.84
standard error	0.01	0.01	0.01	0.01
number of observations	81,671	80,138	86,856	80,844

In the illustrations below we just use a single instrument, an indicator for being born in the first quarter. First let us look at the reduced form regressions of log earnings and years of education on the first quarter of birth dummy:

$$\begin{aligned} \widehat{\text{educ}_i} &= 12.797 - 0.109 \cdot \text{qob}_i \\ &\quad (0.006) \quad (0.013) \end{aligned}$$

and

$$\begin{aligned} \widehat{\log(\text{earnings})}_i &= 5.903 - 0.011 \cdot \text{qob}_i \\ &\quad (0.001) \quad (0.003) \end{aligned}$$

The instrumental variables estimate is the ratio of the reduced form coefficients,

$$\hat{\beta}^{\text{iv}} = \frac{-0.1019}{-0.011} = 0.1020.$$

Now let us interpret this estimate in the context of heterogeneous returns to education, using (16) and (17).. This estimate is an average of returns to education, consisting of two types of averaging. The first averaging is over different levels of education. That is, it is a weighted average of the return to the tenth year of education, to the elevent year of education, and so on.. In addition, for any level, e.g., to moving from nine to ten years of education, it is an average effect where the averaging is over those people whose schooling would have been at least ten years of education if more restrictive compulsory schooling laws had been in effect for them, and who would have had less than ten years of education had they been subject to the looser compulsory schooling laws.

Furthermore, we can estimate how large a fraction of the population is in these categories. First we estimate the

$$\gamma_j = \text{pr}(W_i(1) \geq j > W_i(0) = \text{pr}(W_i \geq j | Z_i = 1) - \text{pr}(W_i \geq j | Z_i = 0)$$

as

$$\hat{\gamma}_j = \frac{1}{N_1} \sum_{i|Z_i=1} 1\{W_i \geq j\} - \frac{1}{N_0} \sum_{i|Z_i=0} 1\{W_i \geq j\}.$$

This gives the unnormalized weight function. We then normalize the weights so they add up to one, $\hat{\lambda}_j = \hat{\gamma}_j / \sum_i \hat{\gamma}_i$.

Figure 1-4 present some of the relevant evidence here. First, Figure 1 gives the distribution of years of education for the Angrist-Krueger data. Figure 2 gives the normalized and Figure 3 gives the unnormalized weight functions. Figure 4 gives the distribution functions of years of education by the two values of the instrument. The most striking feature of these figures (not entirely unanticipated) is that the proportion of individuals in the "complier" subpopulations is extremely small, never more than 2% of the population. This implies that these instrumental variables estimates are averaged only over a very small subpopulation, and that there is little reason to believe that they generalize to the general population. (Nevertheless, this may well be a very interesting subpopulation for some purposes.) The nature

of the instrument also suggests that most of the weight would be just around the number of years that would be required under the compulsory schooling laws. The weight function is actually surprisingly flat, putting weight even on fourteen to fifteen years of education.

References

Abadie, A., (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, Vol 97, No. 457, 284-292.

Abadie, A., (2003), "Semiparametric Instrumental Variable Estimation of Treatment Reponse Models," *Journal of Econometrics*, Vol 113, 231-263.

Angrist, J. D., and G. W. Imbens, (1995), "Two–Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, Vol 90, No. 430, 431-442.

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

Angrist, J. and A. Krueger, (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association* 87, June.

Björklund, A. and R. Moffitt, (1987), "The Estimation of Wage Gains and Welfare Gains in Self–Selection Models", *Review of Economics and Statistics*, Vol. LXIX, 42–49.

Bloom, H., (1984), "Accounting for No–shows in Experimental Evaluation Designs," *Evaluation Review*, 8(2) 225–246.

Chamberlain, G., (1986), "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, Vol 32, 189-218.

Dempster, A., N. Laird, and D. Rubin (1977), "Maximum Likelihood Estimation from Incomplete Data Using the EM Algorithm (with discussion)," *Journal of the Royal*

Statistical Society, Series B, 39, 1-38.

Heckman, J. (1990), "Varieties of Selection Bias," *American Economic Review* Vol 80, Papers and Proceedings, 313-318.

Heckman, J., and R. Robb, (1985), "Alternative Methods for Evaluating the Impact of Interventions,"in Heckman and Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge, Cambridge University Press.

Heckman, J., and E. Vytlacil, (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, Vol. 73(3), 669-738.

Hirano, K., G. Imbens, D. Rubin, and X. Zhou (2000), "Identification and Estimation of Local Average Treatment Effects," *Biostatistics*, Vol. 1(1), 69-88.

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G. W., and D. B. Rubin, (1997), "Bayesian Inference for Causal Effects in Randomized Experiments with Noncompliance,"*Annals of Statistics*, Vol. 25, No. 1, 305–327.

textscManski, C., (2008), *Identification for Prediction and Decision*, Harvard University Press, Cambridge, MA.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Vytlacil, E., (2002), "Independence, Monotonicity, and Latent Index Models: Am Equivalence Result," *Econometrica*, Vol. 70(1), 331-341.

Wooldridge, J. (2001) *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

AEA Lectures **Chicago, IL, January 2012**

Lecture 3, Monday, Jan 9th,pm-pm

Selection on Unobservables:

Part II: Regression Discontinuity Designs

1. Introduction

Since the late 1990s there has been a large number of studies (e.g., Lee, 2001, 2008; VanderKlaauw, 2001) in economics applying and extending Regression Discontinuity (RD) methods from its origins in the statistics literature in the early 60's (Thisthlewaite and Cook, 1960). Here, we review some of the practical issues in implementation of RD methods. The focus is on five specific issues. The first is the importance of graphical analyses as powerful methods for illustrating the design. Second, we suggest using local linear regression methods using only the observations close to the discontinuity point. Third, we discuss choosing the bandwidth using cross validation specifically tailored to the focus on estimation of regression functions on the boundary of the support, following Ludwig and Miller (2005). Fourth, we provide two simple estimators for the asymptotic variance, one of them exploiting the link with instrumental variables methods derived by Hahn, Todd, and VanderKlaauw (2001, HTV). Finally, we discuss a number of specification tests and sensivitity analyses based on tests for (a) discontinuities in the average values for covariates, (b) discontinuities in the conditional density of the forcing variable, as suggested by McCrary (2007), (c) discontinuities in the average outcome at other values of the forcing variable.

2. Sharp and Fuzzy Regression Discontinuity Designs

2.1 Basics

Our discussion will frame the RD design in the context of the modern literature on causal effects and treatment effects, using the potential outcomes framework (Rubin, 1974), rather than the regression framework that was originally used in this literature. For unit i there are two potential outcomes, $Y_i(0)$ and $Y_i(1)$, with the causal effect defined as the difference

$Y_i(1) - Y_i(0)$, and the observed outcome equal to

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases}$$

where $W_i \in \{0, 1\}$ is the binary indicator for the treatment.

The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor (the forcing variable X_i) being on either side of a common threshold. This predictor X_i may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity in the conditional distribution of the outcome, indexed by the value of this covariate at the cutoff value, is interpreted as evidence of a causal effect of the treatment. The design often arises from administrative decisions, where the incentives for units to participate in a program are partly limited for reasons of resource constraints, and clear transparent rules rather than discretion by administrators are used for the allocation of these incentives.

2.2 The Sharp Regression Discontinuity Design

It is useful to distinguish between two designs, the Sharp and the Fuzzy Regression Discontinuity (SRD and FRD from hereon) designs (e.g., Trochim, 1984, 2001; HTV). In the SRD design the assignment W_i is a deterministic function of one of the covariates, the forcing (or treatment-determining) variable X:

$$W_i = 1\{X_i \geq c\}.$$

All units with a covariate value of at least c are in the treatment group (and participation is mandatory for these individuals), and all units with a covariate value less than c are in the control group (members of this group are not eligible for the treatment). In the SRD design we look at the discontinuity in the conditional expectation of the outcome given the covariate to uncover an average causal effect of the treatment:

$$\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1) | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i(0) | X_i = x], \qquad (1)$$

is interpreted as the average causal effect of the treatment at the discontinuity point.

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]. \tag{2}$$

In order to justify this interpretation we make a smoothness assumption. Typically this assumption is formulated in terms of conditional expectations[1]:

Assumption 1 (Continuity of Conditional Regression Functions)

$$\mathbb{E}[Y(0)|X = x] \qquad \text{and} \quad \mathbb{E}[Y(1)|X = x],$$

are continuous in x.

Under this assumption,

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x].$$

The estimand is the difference of two regression functions at a point.

There is a unavoidable need for extrapolation, because by design there are no units with $X_i = c$ for whom we observe $Y_i(0)$. We therefore will exploit the fact that we observe units with covariate values arbitrarily close to c.[2]

As an example of a SRD design, consider the study of the effect of party affiliation of a congressman on congressional voting outcomes by Lee (2007). See also Lee, Moretti and Butler (2004). The key idea is that electoral districts where the share of the vote for a

[1] More generally, one might want to assume that the conditional distribution function is smooth in the covariate. Let $F_{Y(w)|X}(y|x) = \Pr(Y_i(w) \leq y|X_i = x)$ denote the conditional distribution function of $Y_i(w)$ given X_i. Then the general version of the assumption assume that $F_{Y(0)|X}(y|x)$ and $F_{Y(1)|X}(y|x)$ are continuous in x for all y. Both assumptions are stronger than required, as we will only use continuity at $x = c$, but it is rare that it is reasonable to assume continuity for one value of the covariate, but not at other values of the covariate.

[2] Although in principle the first component in the difference in (1) would be straightforward to estimate if we actually observe individuals with $X_i = x$, with continuous covariates we also need to estimate this term by averaging over units with covariate values close to c.

Democrat in a particular election was just under 50% are on average similar in many relevant respects to districts where the share of the Democratic vote was just over 50%, but the small difference in votes leads to an immediate and big difference in the party affiliation of the elected representative. In this case, the party affiliation always jumps at 50%, making this is a SRD design. Lee looks at the incumbency effect. He is interested in the probability of Democrats winning the subsequent election, comparing districts where the Democrats won the previous election with just over 50% of the popular vote with districts where the Democrats lost the previous election with just under 50% of the vote.

2.3 The Fuzzy Regression Discontinuity Design

In the Fuzzy Regression Discontinuity (FRD) design the probability of receiving the treatment need not change from zero to one at the threshold. Instead the design allows for a smaller jump in the probability of assignment to the treatment at the threshold:

$$\lim_{x \downarrow c} \Pr(W_i = 1 | X_i = x) \neq \lim_{x \uparrow c} \Pr(W_i = 1 | X_i = x),$$

without requiring the jump to equal 1. Such a situation can arise if incentives to participate in a program change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation. In this design we interpret the ratio of the jump in the regression of the outcome on the covariate to the jump in the regression of the treatment indicator on the covariate as an average causal effect of the treatment. Formally, the estimand is

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x]}.$$

As an example of a FRD design, consider the study of the effect of financial aid on college attendance by VanderKlaauw (2002). VanderKlaauw looks at the effect of financial aid on acceptance on college admissions. Here X_i is a numerical score assigned to college applicants based on the objective part of the application information (SAT scores, grades) used to streamline the process of assigning financial aid offers. During the initial stages of

the admission process, the applicants are divided into L groups based on discretized values of these scores. Let

$$G_i = \begin{cases} 1 & \text{if } 0 \leq X_i < c_1 \\ 2 & \text{if } c_1 \leq X_i < c_2 \\ \vdots & \\ L & \text{if } c_{L-1} \leq X_i \end{cases}$$

denote the financial aid group. For simplicity, let us focus on the case with $L = 2$, and a single cutoff point c. Having a score just over c will put an applicant in a higher category and increase the chances of financial aid discontinuously compared to having a score just below c. The outcome of interest in the VanderKlaauw study is college attendance. In this case, the statistical association between attendance and the financial aid offer is ambiguous. On the one hand, an aid offer by a college makes that college more attractive to the potential student. This is the causal effect of interest. On the other hand, a student who gets a generous financial aid offer from one college is likely to have better outside opportunities in the form of financial aid offers from other colleges. In the VanderKlaauw application College aid is emphatically not a deterministic function of the financial aid categories, making this a fuzzy RD design. Other components of the college application package that are not incorporated in the numerical score such as the essay and recommendation letters undoubtedly play an important role. Nevertheless, there is a clear discontinuity in the probability of receiving an offer of a larger financial aid package.

Let us first consider the interpretation of τ_{FRD}. HTV exploit the instrumental variables connection to interpret the fuzzy regression discontinuity design when the effect of the treatment varies by unit. Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neigborhood around c. $W_i(x)$ is equal to one if unit i would take or receive the treatment if the cutoff point was equal to x. This requires that the cutoff point is at least in principle manipulable.[3] For example, if X is age, one could imagine changing the age that makes an individual eligible for the treatment from c to $c + \epsilon$. Then it is useful to assume

[3]Alternatively, one could think of the individual characteristic X_i as being manipulable, but in many cases this is an immutable characteristic such as age.

monotonicity (see HTV):

Assumption 2 *$W_i(x)$ is non-increasing in x at $x = c$.*

Next, define compliance status. This concept is similar to that in instrumental variables, e.g., Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996). A complier is a unit such that

$$\lim_{x \downarrow X_i} W_i(x) = 0, \qquad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

Compliers are units who would get the treatment if the cutoff were at X_i or below, but they would not get the treatment if the cutoff were higher than X_i. To be specific, consider an example where individuals with a test score less than c are encouraged for a remedial teaching program (Matsudaira, 2007). Interest is in the effect of the remedial teaching program on subsequent test scores. Compliers are individuals who would participate if encouraged (if the cutoff for encouragement is below or equal to their actual score), but not if not encouraged (if the cutoff for encouragement is higher than their actual score). Then

$$\frac{\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i | X_i = x]}$$

$$= \mathbb{E}[Y_i(1) - Y_i(0) | \text{unit } i \text{ is a complier and } X_i = c].$$

The estimand is an average effect of the treatment, but only averaged for units with $X_i = c$ (by regression discontinuity), and only for compliers (people who are affected by the threshold).

3. The FRD Design, Unconfoundedness and External Validity

3.1 The FRD Design and Unconfoundedness

In the FRD setting it is useful to contrast the RD approach with estimation of average causal effects under unconfoundedness. The unconfoundedness assumption, e.g., Rosenbaum

and Rubin (1983), Imbens (2004), requires that

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \,\Big|\, X_i.$$

If this assumption holds, then we can estimate the average effect of the treatment at $X_i = c$ as

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i|W_i = 1, X_i = c] - \mathbb{E}[Y_i|W_i = 0, X_i = c].$$

This approach does not exploit the jump in the probability of assignment at the discontinuity point. Instead it assumes that differences between treated and control units with $X_i = c$ are interpretable as average causal effects.

In contrast, the assumptions underlying an FRD analysis implies that comparing treated and control units with $X_i = c$ is likely to be the wrong approach. Treated units with $X_i = c$ include compliers and alwaystakers, and control units at $X_i = c$ consist only of nevertakers. Comparing these different types of units has no causal interpretation under the FRD assumptions. Although, in principle, one cannot test the unconfoundedness assumption, one aspect of the problem makes this assumption fairly implausible. Unconfoundedness is fundamentally based on units being comparable if their covariates are similar. This is not an attractive assumption in the current setting where the probability of receiving the treatment is discontinuous in the covariate. Thus units with similar values of the forcing variable (but on different sides of the threshold) must be different in some important way related to the receipt of treatment. Unless there is a substantive argument that this difference is immaterial for the comparison of the outcomes of interest, an analysis based on unconfoundedness is not attractive in this setting. Moroever, even if unconfoundedness holds, if the expected values of the potential outcomes given the forcing variables are continuous in the forcing variable, and the monotonicity assumption holds, then the FRD approach will still estimate a well defined average causal effect, equal to $\mathbb{E}[Y_i(1) - Y_i(0|X_i = c]$ under unconfoundedness. One can estimate this effect more efficiently under unconfoundedness, but the FRD approach remains consistent for this average effect.

3.2 The FRD Design and External Validity

One important aspect of both the SRD and FRD designs is that they at best provide estimates of the average effect for a subpopulation, namely the subpopulation with covariate value equal to $X_i = c$. The FRD design restricts the relevant subpopulation even further to that of compliers at this value of the covariate. Without strong assumptions justifying extrapolation to other subpopulations (e.g., homogeneity of the treatment effect) the designs never allow the researcher to estimate the overall (population) average effect of the treatment. In that sense the design has fundamentally only a limited degree of external validity, although the specific average effect that is identified may well be of special interest, for example in cases where the policy question concerns changing the location of the threshold. The advantage of RD designs compared to other non-experimental analyses that may have more external validity such as those based on unconfoundedness, is that RD designs generally have a relatively high degree of internal validity in settings where they are applicable.

4. Graphical Analyses

4.1 Introduction

Graphical analyses should be an integral part of any RD analysis. The nature of RD designs suggests that the effect of the treatment of interest can be measured by the value of the discontinuity in the expected value of the outcome at a particular point. Inspecting the estimated version of this conditional expectation is a simple yet powerful way to visualize the identification strategy. Moreover, to assess the credibility of the RD strategy, it is useful to inspect two additional graphs. The estimators we discuss later use more sophisticated methods for smoothing but these basic plots will convey much of the intuition. For strikingly clear examples of such plots, see Lee, Moretti, and Butler (2004), Lalive (2007), and Lee (2007). Two figures from Lee (2007) are attached.

4.2 Outcomes by Forcing Variable

The first plot is a histogram-type estimate of the average value of the outcome by the forcing variable. For some binwidth h, and for some number of bins K_0 and K_1 to the left and

right of the cutoff value, respectively, construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1) \cdot h.$$

Then calculate the number of observations in each bin,

$$N_k = \sum_{i=1}^{N} 1\{b_k < X_i \leq b_{k+1}\},$$

and the average outcome in the bin:

$$\overline{Y}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The first plot of interest is that of the $\overline{Y}_k$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k = (b_k + b_{k+1})/2$. The question is whether around the threshold c there is any evidence of a jump in the conditional mean of the outcome. The formal statistical analyses discussed below are essentially just sophisticated versions of this, and if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes. In addition to inspecting whether there is a jump at this value of the covariate, one should inspect the graph to see whether there are any other jumps in the conditional expectation of Y_i given X_i that are comparable to, or larger than, the discontinuity at the cutoff value. If so, and if one cannot explain such jumps on substantive grounds, it would call into question the interpretation of the jump at the threshold as the causal effect of the treatment. In order to optimize the visual clarity it is important to calculate averages that are not smoothed over the cutoff point. The attached figure is taken from the paper by Lee (2007).

4.2 Covariates by Forcing Variable

The second set of plots compares average values of other covariates in the K bins. Specifically, let Z_i be the M-vector of additional covariates, with m-th element Z_{im}. Then calculate

$$\overline{Z}_{km} = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Z_{im} \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The second plot of interest is that of the $\overline{Z}_{km}$, for $k = 1, K$ against the mid point of the bins, $\tilde{b}_k$, for all $m = 1, \ldots, M$. Lee (2007) presents such a figure for a lagged value of the outcome, namely the election results from a prior election, against the vote share in the last election. In the case of FRD designs, it is also particularly useful to plot the mean values of the treatment variable W_i to make sure there is indeed a jump in the probability of treatment at the cutoff point. Plotting other covariates is also useful for detecting possible specification problems (see Section 8) in the case of either SRD or FRD designs.

4.3 The Density of the Forcing Variable

In the third graph one should plot the number of observations in each bin, N_k, against the mid points $\tilde{b}_k$. This plot can be used to inspect whether there is a discontinuity in the distribution of the forcing variable X at the threshold. McCrary (2007) suggests that such discontinuity would raise the question whether the value of this covariate was manipulated by the individual agents, invalidating the design. For example, suppose that the forcing variable is a test score. If individuals know the threshold and have the option of re-taking the test, individuals with test scores just below the threshold may do so, and invalidate the design. Such a situation would lead to a discontinuity of the conditional density of the test score at the threshold, and thus be detectable in plots such as described here. See Section 8 for more discussion of the specification tests based on this idea.

5. Estimation: Local Linear Regression

5.1 Nonparametric Regression at the Boundary

The practical estimation of the treatment effect τ in both the SRD and FRD designs is largely standard nonparametric regression (e.g., Pagan and Ullah, 1999; Härdle, 1990;

Li and Racine, 2007). However, there are two unusual features to estimation in the RD setting. First, we are interested in the regression function at a single point, and second, that single point is a boundary point. As a result, standard nonparametric kernel regression does not work very well. At boundary points, such estimators have a slower rate of convergence than they do at interior points. Standard methods for choosing the bandwidth are also not designed to provide good choices in this setting.

5.2 Local Linear Regression

Here we discuss local linear regression (Fan and Gijbels, 1996). Instead of locally fitting a constant function, we can fit linear regression functions to the observations within a distance h on either side of the discontinuity point:

$$\min_{\alpha_l,\beta_l} \sum_{i|c-h<X_i<c}^{N} \left(Y_i - \alpha_l - \beta_l \cdot (X_i - c)\right)^2,$$

and

$$\min_{\alpha_r,\beta_r} \sum_{i|c\leq X_i<c+h}^{N} \left(Y_i - \alpha_r - \beta_r \cdot (X_i - c)\right)^2.$$

The value of $\mu_l(c)$ and $\mu_r(c)$ are then estimated as

$$\widehat{\mu_l(c)} = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l, \qquad \text{and} \quad \widehat{\mu_r(c)} = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r,$$

Given these estimates, the average treatment effect is estimated as

$$\hat{\tau}_{\text{SRD}} = \hat{\alpha}_r - \hat{\alpha}_l.$$

Alternatively one can estimate the average effect directly in a single regression, by solving

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} 1\{c - h \leq X_i \leq c + h\} \cdot \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i\right)^2,$$

which will numerically yield the same estimate of τ_{SRD}.

We can make the nonparametric regression more sophisticated by using weights that decrease smoothly as the distance to the cutoff point increases, instead of the zero/one weights based on the rectangular kernel. However, even in this simple case the asymptotic bias can be shown to be of order h^2, and the more sophisticated kernels rarely make much difference. Furthermore, if using different weights from a more sophisticated kernel does make a difference, it likely suggests that the results are highly sensitive to the choice of bandwidth. So the only case where more sophisticated kernels may make a difference is when the estimates are not very credible anyway because of too much sensitivity to the choice of bandwidth. From a practical point of view one may just focus on the simple rectangular kernel, but verify the robustness of the results to different choices of bandwidth.

For inference we can use standard least squares methods. Under appropriate conditions on the rate at which the bandwidth goes to zero as the sample size increases, the resulting estimates will be asymptotically normally distributed, and the (robust) standard errors from least squares theory will be justified. Using the results from HTV, the optimal bandwidth is $h \propto N^{-1/5}$. Under this sequence of bandwidths the asymptotic distribution of the estimator $\hat{\tau}$ will have a non-zero bias. If one does some undersmoothing, by requiring that $h \propto N^{-\delta}$ with $1/5 < \delta < 2/5$, then the asymptotic bias disappears and standard least squares variance estimators will lead to valid confidence intervals.

5.3 Covariates

Often there are additional covariates available in addition to the forcing covariate that is the basis of the assignment mechanism. These covariates can be used to eliminate small sample biases present in the basic specification, and improve the precision. In addition, they can be useful for evaluating the plausibility of the identification strategy, as discussed in Section 8.1. Let the additional vector of covariates be denoted by Z_i. We make three observations on the role of these additional covariates.

The first and most important point is that the presence of these covariates rarely changes the identification strategy. Typically, the conditional distribution of the covariates Z given X

is continuous at $x = c$. If such discontinuities in other covariates are found, the justification of the identification strategy may be questionable. If the conditional distribution of Z given X is continuous at $x = c$, then including Z in the regression

$$\min_{\alpha,\beta,\tau,\delta} \sum_{i=1}^{N} 1\{c-h \leq X_i \leq c+h\} \cdot (Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i - \delta' Z_i)^2 ,$$

will have little effect on the expected value of the estimator for τ, since conditional on X being close to c, the additional covariates Z are independent of W.

The second point is that even though with X very close to c, the presence of Z in the regression does not affect any bias, in practice we often include observations with values of X not too close to c. In that case, including additional covariates may eliminate some bias that is the result of the inclusion of these additional observations.

Third, the presence of the covariates can improve precision if Z is correlated with the potential outcomes. This is the standard argument, which also supports the inclusion of covariates even in analyses of randomized experiments. In practice the variance reduction will be relatively small unless the contribution to the $\mathbb{R}^2$ from the additional regressors is substantial.

5.4 Estimation for the Fuzzy Regression Discontinuity Design

In the FRD design, we need to estimate the ratio of two differences. The estimation issues we discussed earlier in the case of the SRD arise now for both differences. In particular, there are substantial biases if we do simple kernel regressions. Instead, it is again likely to be better to use local linear regression. We use a uniform (rectangular) kernel, with the same bandwidth for estimation of the discontinuity in the outcome and treatment regressions.

First, consider local linear regression for the outcome, on both sides of the discontinuity point. Let

$$\left(\hat{\alpha}_{yl}, \hat{\beta}_{yl}\right) = \arg\min_{\alpha_{yl},\beta_{yl}} \sum_{i:c-h\leq X_i<c} \left(Y_i - \alpha_{yl} - \beta_{yl} \cdot (X_i - c)\right)^2 , \tag{3}$$

$$\left(\hat{\alpha}_{yr}, \hat{\beta}_{yr}\right) = \arg\min_{\alpha_{yr},\beta_{yr}} \sum_{i:c\leq X_i\leq c+h} \left(Y_i - \alpha_{yr} - \beta_{yr}\cdot(X_i - c)\right)^2. \tag{4}$$

The magnitude of the discontinuity in the outcome regression is then estimated as $\hat{\tau}_y = \hat{\alpha}_{yr} - \hat{\alpha}_{yl}$. Second, consider the two local linear regression for the treatment indicator:

$$\left(\hat{\alpha}_{wl}, \hat{\beta}_{wl}\right) = \arg\min_{\alpha_{wl},\beta_{wl}} \sum_{i:c-h\leq X_i<c} \left(W_i - \alpha_{wl} - \beta_{wl}\cdot(X_i - c)\right)^2, \tag{5}$$

$$\left(\hat{\alpha}_{wr}, \hat{\beta}_{wr}\right) = \arg\min_{\alpha_{wr},\beta_{wr}} \sum_{i:c\leq X_i\leq c+h} \left(Y_i - \alpha_{wr} - \beta_{wr}\cdot(X_i - c)\right)^2. \tag{6}$$

The magnitude of the discontinuity in the treatment regression is then estimated as $\hat{\tau}_w = \hat{\alpha}_{wr} - \hat{\alpha}_{wl}$. Finally, we estimate the effect of interest as the ratio of the two discontinuities:

$$\hat{\tau}_{\text{FRD}} = \frac{\hat{\tau}_y}{\hat{\tau}_w} = \frac{\hat{\alpha}_{yr} - \hat{\alpha}_{yl}}{\hat{\alpha}_{wr} - \hat{\alpha}_{wl}}. \tag{7}$$

Because of the specific implementation we use here, with a uniform kernel, and the same bandwidth for estimation of the denominator and the numerator, we can characterize the estimator for τ as a Two-Stage-Least-Squares (TSLS) estimator (See HTV). This equality still holds when we use local linear regression and include additional regressors. Define

$$V_i = \begin{pmatrix} 1 \\ 1\{X_i < c\}\cdot(X_i - c) \\ 1\{X_i \geq c\}\cdot(X_i - c) \end{pmatrix}, \qquad \text{and} \quad \delta = \begin{pmatrix} \alpha_{yl} \\ \beta_{yl} \\ \beta_{yr} \end{pmatrix}. \tag{8}$$

Then we can write

$$Y_i = \delta' V_i + \tau\cdot W_i + \varepsilon_i. \tag{9}$$

Estimating τ based on the regression function (9) by TSLS methods, with the indicator $1\{X_i \geq c\}$ as the excluded instrument and V_i as the set of exogenous variables is numerically identical to $\hat{\tau}_{\text{FRD}}$ as given in (7).

6. Bandwidth Selection

An important issue in practice is the selection of the smoothing parameter, the binwidth h. Most of the empirical literature uses bandwidth selection methods that are borrowed from the general nonparametric estimation literature, without adjusting for the special features of the RD design. One exception is the work by by Ludwig and Miller (2005, 2007) that develops cross-validation methods. See also Imbens and Lemieux (2008).

Here we focus on a recent alternative, a plug in method developed by Imbens and Kalyanaraman (2008). Initially we focus on the SRD case, and in Section 6.2 we extend the recommendations to the FRD setting.

6.1 Bandwidth Selection for the SRD Design

To set up the bandwidth choice problem we generalize the notation slightly. In the SRD setting we are interested in the

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mu(x) - \lim_{x \uparrow c} \mu(x),$$

where $\mu(x) = \mathbb{E}[Y_i | X_i = x]$. We estimate the two terms as

$$\hat{\mu}_r(c) = \widehat{\lim_{x \downarrow c} \mu(x)} = \hat{\alpha}_r(c), \qquad \text{and} \quad \hat{\mu}_l(c) = \widehat{\lim_{x \uparrow c} \mu(x)} = \hat{\alpha}_l(c),$$

where $\hat{\alpha}_l(x)$ and $\hat{\beta}_l(x)$ solve

$$\left(\hat{\alpha}_l(x), \hat{\beta}_l(x)\right) = \arg\min_{\alpha,\beta} \sum_{j | x-h<X_j<x} \left(Y_j - \alpha - \beta \cdot (X_j - x)\right)^2. \tag{10}$$

and $\hat{\alpha}_r(x)$ and $\hat{\beta}_r(x)$ solve

$$\left(\hat{\alpha}_r(x), \hat{\beta}_r(x)\right) = \arg\min_{\alpha,\beta} \sum_{j | x<X_j<x+h} \left(Y_j - \alpha - \beta \cdot (X_j - x)\right)^2. \tag{11}$$

Imbens and Kalyanaraman focus on minimizing

$$\mathbb{E}\left[\left((\hat{\mu}_r(c) - \hat{\mu}_l(c)) - (\mu_r(c) - \mu_l(c))\right)^2\right]. \tag{12}$$

Both $\hat{\mu}_l(c)$ and $\hat{\mu}_r(c)$ are based on local linear estimators, with the same bandwidth h. This is not necessary, but in practice it is unlikely that the differences in the optimal bandwidth are substantial enough that they can be exploited with typical sample sizes. Under homoskedasticity, so that the variance on the right and the left are the same, $\sigma^2(c) = \sigma_l^2(c) = \sigma_r^2(c)$, and continuity of the density at c, $f_X(c) = f_{X,l}(c) = f_{X,r}(c)$, the leading terms in the expected squared error are

$$\mathbb{E}\left[\left((\hat{\mu}_r(c) - \hat{\mu}_l(c)) - (\mu_r(c) - \mu_l(c))\right)^2\right].$$

$$= h^4 \cdot \left(\left(\frac{\partial^2 m_r}{\partial x^2}(c) - \frac{\partial^2 m_l}{\partial x^2}(c)\right)^2\right) \cdot C_1 + \frac{C_2}{N \cdot h} \cdot \left(\frac{2 \cdot \sigma^2(c)}{f_X(c)}\right) + o\left(h^4 + \frac{1}{N \cdot h}\right)$$

Here C_1 and C_2 are constants that depend on the kernel:

$$C_1 = \frac{1}{4} \cdot \left(\frac{\nu_2^2 - \nu_1\nu_3}{\nu_2\nu_0 - \nu_1^2}\right)^2 \qquad C_2 = \frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_2\nu_0 - \nu_1^2)^2}$$

$$\nu_j = \int_0^\infty u^j K(u)du, \qquad \text{and} \quad \pi_j = \int_0^\infty u^j K^2(u)du.$$

Under the same conditions the optimal bandwidth is

$$h_{\text{opt}} = \left(\frac{C_2}{C_1}\right)^{1/5} \cdot \left(\frac{2 \cdot \sigma^2(c)}{f_X(c)}\right)^{1/5} \left(\frac{1}{\left(\frac{\partial^2 m_r}{\partial x^2}(c) - \frac{\partial^2 m_l}{\partial x^2}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}. \tag{13}$$

Imbens and Kalyanaraman (2008) describe a data-dependent algorithm for estimating the optimal bandwidth. The main difficulty is in estimating the second derivatives of the regression function, and suggest a modification to avoid the denominator getting too close to zero when the second derivatives are estimated imprecisely.

6.2 Bandwidth Selection for the FRD Design

In the FRD design, there are four regression functions that need to be estimated: the expected outcome given the forcing variable, both on the left and right of the cutoff point,

and the expected value of the treatment, again on the left and right of the cutoff point. In principle, we can use different binwidths for each of the four nonparametric regressions.

In the section on the SRD design, we argued in favor of using identical bandwidths for the regressions on both sides of the cutoff point. The argument is not so clear for the pairs of regressions functions by outcome we have here, and so in principle we have two optimal bandwidths, each based on minimizing a criterion like (13). It is likely that the conditional expectation of the treatment is relatively flat compared to the conditional expectation of the outcome variable, suggesting one should use a larger binwidth for estimating the former.[4] Nevertheless, in practice it is appealing to use the same binwidth for numerator and denominator. Since typically the size of the discontinuity is much more marked in the expected value of the treatment, one option is to use the optimal bandwidth based on the outcome discontinuity.

[4]In the extreme case of the SRD design the conditional expectation of W given X is flat on both sides of the threshold, and so the optimal bandwidth would be infinity. Therefore, in practice it is likely that the optimal bandwidth would be larger for estimating the jump in the conditional expectation of the treatment than in estimating the jump in the conditional expectation of the outcome.

7. Inference

We now discuss some asymptotic properties for the estimator for the FRD case given in (7) or its alternative representation in (9).[5] More general results are given in HTV. We continue to make some simplifying assumptions. First, as in the previous sections, we use a uniform kernel. Second, we use the same bandwidth for the estimator for the jump in the conditional expectations of the outcome and treatment. Third, we undersmooth, so that the square of the bias vanishes faster than the variance, and we can ignore the bias in the construction of confidence intervals. Fourth, we continue to use the local linear estimator. Under these assumptions we give an explicit expression for the asymptotic variance, and present two estimators for the asymptotic variance. The first estimator follows explicitly the analytic form for the asymptotic variance, and substitutes estimates for the unknown quantities. The second estimator is the standard robust variance for the Two-Stage-Least-Squares (TSLS) estimator, based on the sample obtained by discarding observations when the forcing covariate is more than h away from the cutoff point. Both are robust to heteroskedasticity.

7.1 The Asymptotic Variance

To characterize the asymptotic variance we need a couple of additional pieces of notation. Define the four variances

$$\sigma^2_{Yl} = \lim_{x\uparrow c} \mathrm{Var}(Y_i|X_i = x), \qquad \sigma^2_{Yr} = \lim_{x\downarrow c} \mathrm{Var}(Y_i|X_i = x),$$

$$\sigma^2_{Wl} = \lim_{x\uparrow c} \mathrm{Var}(W_i|X_i = x), \qquad \sigma^2_{Wr} = \lim_{x\downarrow c} \mathrm{Var}(W_i|X_i = x),$$

and the two covariances

$$C_{YWl} = \lim_{x\uparrow c} \mathrm{Cov}(Y_i, W_i|X_i = x), \qquad C_{YWr} = \lim_{x\downarrow c} \mathrm{Cov}(Y_i, W_i|X_i = x).$$

Note that because of the binary nature of W, it follows that $\sigma^2_{Wl} = \mu_{Wl} \cdot (1 - \mu_{Wl})$, where $\mu_{Wl} = \lim_{x\uparrow c} \Pr(W_i = 1|X_i = x)$, and similarly for σ^2_{Wr}. To discuss the asymptotic variance

[5] The results for the SRD design are a special case of these for the FRD design.

of $\hat{\tau}$ it is useful to break it up in three pieces. The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ is

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl}\right). \tag{14}$$

The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl}\right) \tag{15}$$

The asymptotic covariance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ and $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$C_{\tau_y,\tau_w} = \frac{4}{f_X(c)} \cdot \left(C_{YWr} + C_{YWl}\right). \tag{16}$$

Finally, the asymptotic distribution has the form

$$\sqrt{Nh} \cdot (\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\tau_w^2} \cdot V_{\tau_y} + \frac{\tau_y^2}{\tau_w^4} \cdot V_{\tau_w} - 2 \cdot \frac{\tau_y}{\tau_w^3} \cdot C_{\tau_y,\tau_w}\right). \tag{17}$$

This asymptotic distribution is a special case of that in HTV (page 208), using the rectangular kernel, and with $h = N^{-\delta}$, for $1/5 < \delta < 2/5$ (so that the asymptotic bias can be ignored).

7.2 A Plug-in Estimator for the Asymptotic Variance

We now discuss two estimators for the asymptotic variance of $\hat{\tau}$. First, we can estimate the asymptotic variance of $\hat{\tau}$ by estimating each of the components, τ_w, τ_y, V_{τ_w}, V_{τ_y}, and C_{τ_y,τ_w} and substituting them into the expression for the variance in (17). In order to do this we first estimate the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{\mu}_y(X_i) = Y_i - 1\{X_i < c\} \cdot \hat{\alpha}_{yl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{yr},$$

$$\hat{\eta}_i = W_i - \hat{\mu}_w(X_i) = W_i - 1\{X_i < c\} \cdot \hat{\alpha}_{wl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{wr}.$$

Then we estimate the variances and covariances consistently as

$$\hat{\sigma}^2_{Yl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}_i^2, \qquad \hat{\sigma}^2_{Yr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}_i^2,$$

$$\hat{\sigma}^2_{Wl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\eta}_i^2, \qquad \hat{\sigma}^2_{Wr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\eta}_i^2,$$

$$\hat{C}_{YWl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}_i \cdot \hat{\eta}_i, \qquad \hat{C}_{YWr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}_i \cdot \hat{\eta}_i.$$

Finally, we estimate the density consistently as

$$\hat{f}_X(x) = \frac{N_{hl} + N_{hr}}{2 \cdot N \cdot h}.$$

Then we can plug in the estimated components of V_{τ_y}, V_{τ_W}, and C_{τ_Y,τ_W} from (14)-(16), and finally substitute these into the variance expression in (17).

7.3 The TSLS Variance Estimator

The second estimator for the asymptotic variance of $\hat{\tau}$ exploits the interpretation of the $\hat{\tau}$ as a TSLS estimator, given in (9). The variance estimator is equal to the robust variance for TSLS based on the subsample of observations with $c - h \leq X_i \leq c + h$, using the indicator $1\{X_i \geq c\}$ as the excluded instrument, the treatment W_i as the endogenous regressor and the V_i defined in (8) as the exogenous covariates.

8. Specification Testing

There are generally two main conceptual concerns in the application of RD designs, sharp or fuzzy. A first concern about RD designs is the possibility of other changes at the same cutoff value of the covariate. Such changes may affect the outcome, and these effects may be attributed erroneously to the treatment of interest. The second concern is that of manipulation of the covariate value.

8.1 Tests Involving Covariates

One category of tests involves testing the null hypothesis of a zero average effect on pseudo outcomes known not to be affected by the treatment. Such variables includes covariates that are by definition not affected by the treatment. Such tests are familiar from settings with

identification based on unconfoundedness assumptions. In most cases, the reason for the discontinuity in the probability of the treatment does not suggest a discontinuity in the average value of covariates. If we find such a discontinuity, it typically casts doubt on the assumptions underlying the RD design. See the second part of the Lee (2007) figure for an example.

8.2 Tests of Continuity of the Density

The second test is conceptually somewhat different, and unique to the RD setting. McCrary (2007) suggests testing the null hypothesis of continuity of the density of the covariate that underlies the assignment at the discontinuity point, against the alternative of a jump in the density function at that point. Again, in principle, one does not need continuity of the density of X at c, but a discontinuity is suggestive of violations of the no-manipulation assumption. If in fact individuals partly manage to manipulate the value of X in order to be on one side of the boundary rather than the other, one might expect to see a discontinuity in this density at the discontinuity point.

8.3 Testing for Jumps at Non-discontinuity Points

Taking the subsample with $X_i < c$ we can test for a jump in the conditional mean of the outcome at the median of the forcing variable. To implement the test, use the same method for selecting the binwidth as before. Also estimate the standard errors of the jump and use this to test the hypothesis of a zero jump. Repeat this using the subsample to the right of the cutoff point with $X_i \geq c$. Now estimate the jump in the regression function and at $q_{X,1/2,r}$, and test whether it is equal to zero.

8.4 RD Designs with Misspecification

Lee and Card (2007) study the case where the forcing variable variable X is discrete. In practice this is of course always true. This implies that ultimately one relies for identification on functional form assumptions for the regression function $\mu(x)$. Lee and Card consider a parametric specification for the regression function that does not fully saturate the model, that is, it has fewer free parameters than there are support points. They then interpret the

deviation between the true conditional expectation $\mathbb{E}[Y|X = x]$ and the estimated regression function as random specification error that introduces a group structure on the standard errors. Lee and Card then show how to incorporate this group structure into the standard errors for the estimated treatment effect. Within the local linear regression framework discussed in the current paper one can calculate the Lee-Card standard errors (possibly based on slightly coarsened covariate data if X is close to continuous) and compare them to the conventional ones.

8.5 Sensitivity to the Choice of Bandwidth

One should investigate the sensitivity of the inferences to this choice, for example, by including results for bandwidths twice (or four times) and half (or a quarter of) the size of the originally chosen bandwidth. Obviously, such bandwidth choices affect both estimates and standard errors, but if the results are critically dependent on a particular bandwidth choice, they are clearly less credible than if they are robust to such variation in bandwidths.

8.6 Comparisons to Estimates Based on Unconfoundedness in the FRD Design

If we have an FRD design, we can also consider estimates based on unconfoundedness. Inspecting such estimates and especially their variation over the range of the covariate can be useful. If we find that for a range of values of X, our estimate of the average effect of the treatment is relatively constant and similar to that based on the FRD approach, one would be more confident in both sets of estimates.

9. Illustration Based on Lee Election Data

Here we illustrate some of the methods discussed in these notes using data from Lee's paper on the effect of incumbency in congressional elections. The forcing variable in Lee's study is the difference in the vote share of the Democratic part versus the Republican party in the last election. The threshold is zero: if the difference is greater than zero the Democrats won the last election, and if not the Republicans won.

We consider two outcomes. The first is an indicator for democrats winning the next election, and the second is the vote share for the democrats in next election, . We also use one covariate, the vote share for the Democrats in the preceeding election. Lee's data set consists of 6558 congressional elections.

We use a uniform kernel with support $[-0.5, 0.5]$. We calculate the optimal bandwidth based on the Imbens-Kalyanaraman procedure. Table 1 presents the results.

Figures 1-3 present histogram-based estimates of the regression functions, and Figure 4 a histogram estimate of the density of the forcing variable. The binwidth in these figures is

Outcome	IK Bandwidth	Estimate	(s.e.)
Dem Win Next Elect	0.36	0.082	(0.010)
Demt Margin Next Election	0.27	0.412	(0.039)
Dem Margin Prev Election	0.28	-0.003	(0.013)

$h = 0.05$. There is no evidence of a discontinuity in the density function, or in the expected value of the covariate, both supportive of the RD approach to this problem.

References

Angrist, J.D., G.W. Imbens and D.B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-472.

Angrist, J.D. and A.B. Krueger, (1991), Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics* 106, 979-1014.

Angrist, J.D., and V. Lavy, (1999), Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics* 114, 533-575.

Black, S., (1999), Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics* 114, 577-599.

Card, D., A. Mas, and J. Rothstein, (2006), Tipping and the Dynamics of Segregation in Neighborhoods and Schools, Unpublished Manuscript, Department of Economics, Princeton University.

Chay, K., and M. Greenstone, (2005), Does Air Quality Matter; Evidence from the Housing Market, *Journal of Political Economy* 113, 376-424.

Cook, T., (2007), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, forthcoming, *Journal of Econometrics*.

DiNardo, J., and D.S. Lee, (2004), Economic Impacts of New Unionization on Private Sector Employers: 1984-2001, *Quarterly Journal of Economics* 119, 1383-1441.

Fan, J. and I. Gijbels, (1996), *Local Polynomial Modelling and Its Applications* (Chapman and Hall, London).

Gerber, A., D. Kessler, and M. Meredith, (2008), The Persuasive Effects of Direct Maio: A Regression-Discontinuity Approach, NBER Working Paper 14206.

Hahn, J., P. Todd and W. Van der Klaauw, (2001), Identification and Estimation of Treatment Effects with a Regression Discontinuity Design, *Econometrica* 69, 201-209.

Härdle, W., (1990), *Applied Nonparametric Regression* (Cambridge University Press, New York).

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., and K. Kalyanaraman, (2008), "Optimal Bandwidth Selection in Regression Discontinuity Designs,"unpublished manuscript, Department of Economics, Harvard University.

Imbens, G., and T. Lemieux, (2008) "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*.

Lee, D, (2001), "The Electoral Advantage of Incumbency and the Voter's Valuation of Political Experience: A Regression Discontinuity Analysis of Close Elections,"unpublished manuscript, Department of Economics, University of California.

Lee, D.S., (2008), "Randomized Experiments from Non-random Selection in U.S. House Elections", *Journal of Econometrics*,Vol 142(2): 675-697.

Lee, D.S. and D. Card, (2007), Regression Discontinuity Inference with Specification Error, forthcoming, *Journal of Econometrics*.

Lee, D.S., Moretti, E., and M. Butler, (2004), Do Voters Affect or Elect Policies? Evidence from the U.S. House, *Quarterly Journal of Economics* 119, 807-859.

Lemieux, T. and K. Milligan, (2007), Incentive Effects of Social Assistance: A Regression Discontinuity Approach, forthcoming, *Journal of Econometrics*.

Ludwig, J., and D. Miller, (2005), Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, NBER working paper 11702.

McCrary, J., (2007), Testing for Manipulation of the Running Variable in the Regression Discontinuity Design, forthcoming, *Journal of Econometrics*.

McEwan, P., and J. Shapiro, (2007),The Benefits of Delayed Primary School En-

rollment: Discontinuity Estimates using exact Birth Dates," Unpublished manuscript.

Pagan, A. and A. Ullah, (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Porter, J., (2003), Estimation in the Regression Discontinuity Model," mimeo, Department of Economics, University of Wisconsin, http://www.ssc.wisc.edu/ jporter/reg_discont_2003.pdf.

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston).

Thistlewaite, D., and D. Campbell, (1960), Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment, *Journal of Educational Psychology* 51, 309-317.

Trochim, W., (1984), *Research Design for Program Evaluation; The Regression-discontinuity Design* (Sage Publications, Beverly Hills, CA).

Trochim, W., (2001), Regression-Discontinuity Design, in N.J. Smelser and P.B Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* 19 (Elsevier North-Holland, Amsterdam) 12940-12945.

Van Der Klaauw, W., (2002), Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–discontinuity Approach, *International Economic Review* 43, 1249-1287.

Fig 1: Regression Function for Margin

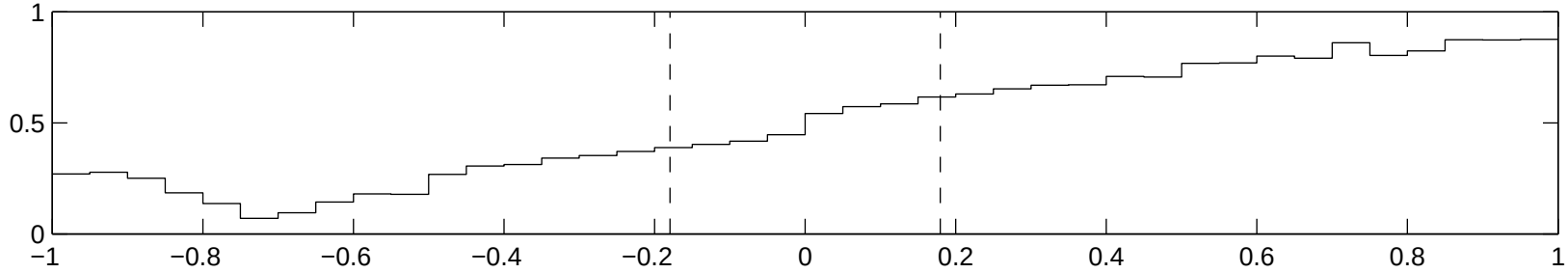


Fig 2: Regression Function for Winning

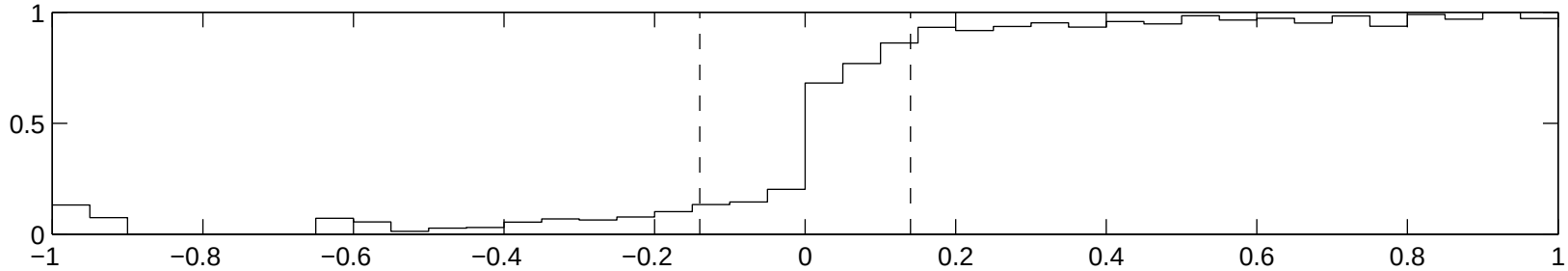


Fig 3: Regression Function for Covariate

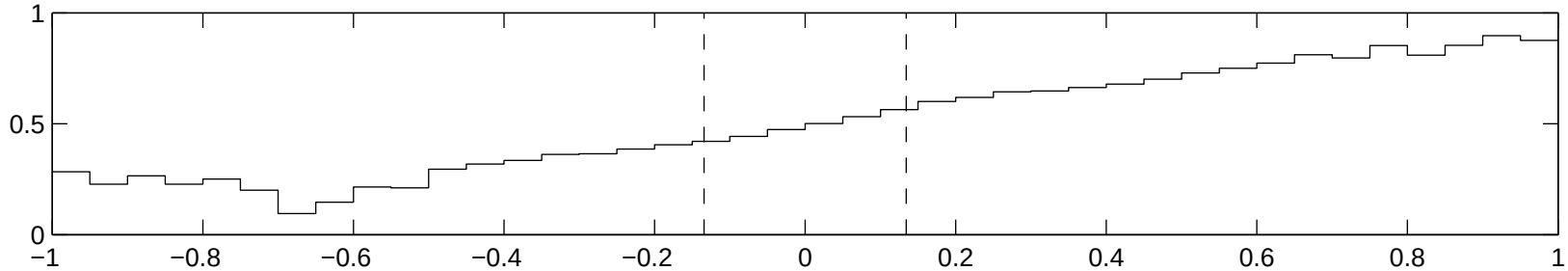
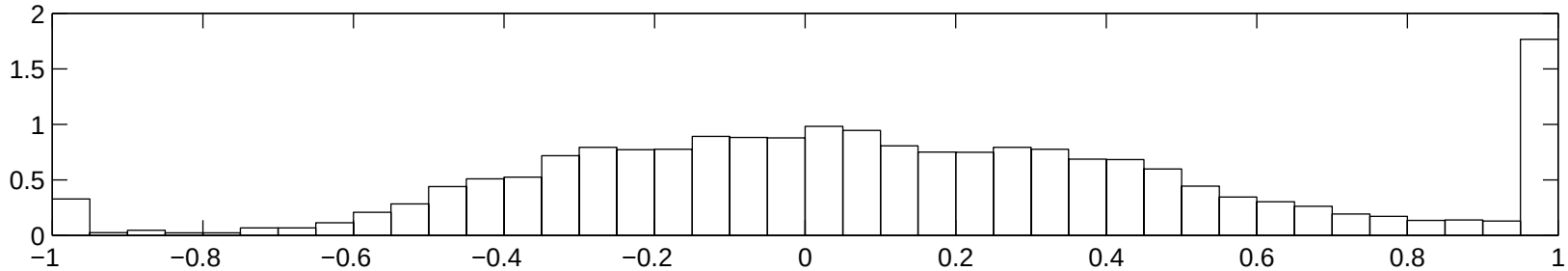


Fig 4: Density for Forcing Variable



"Cross-Section Econometrics"

Lecture 3

Selection on Unobservables:

Instrumental Variables and Regression Discontinuity

Guido Imbens

AEA Lectures, Chicago, January 2012

Part I: Instrumental Variables and Local Average Treatment Effects

1. Introduction
2. Basics
3. Local Average Treatment Effects
4. Extrapolation to the Population
5. Illustration

1. Introduction

1. Instrumental variables estimate average treatment effects, with the average depending on the instruments.
2. Population averages are only estimable under unrealistically strong assumptions ("identification at infinity", or under the constant effect).
3. Compliers (for whom we can identify effects) are not necessarily the subpopulations that are *ex ante* the most interesting subpopulations, but need extrapolation for others.
4. The set up here allows the researcher to sharply separate the extrapolation to the (sub-)population of interest from exploration of the information in the data.

2. Basics

Linear IV with Constant Coefficients. Standard set up:

$$Y_i = \beta_0 + \beta_1 \cdot W_i + \varepsilon_i.$$

There is concern that the regressor W_i is endogenous, correlated with ε_i. Suppose that we have an instrument Z_i that is both uncorrelated with ε_i and correlated with W_i.

In the single instrument / single endogenous regressor, we end up with the ratio of covariances

$$\hat{\beta}_1^{\rm IV} = \frac{\frac{1}{N}\sum_{i=1}^N (Y_i - \bar{Y}) \cdot (Z_i - \bar{Z})}{\frac{1}{N}\sum_{i=1}^N (W_i - \bar{W}) \cdot (Z_i - \bar{Z})}.$$

Using a central limit theorem for all the moments and the delta method we can infer the large sample distribution without additional assumptions.

Potential Outcome Set Up

Let $Y_i(0)$ and $Y_i(1)$ be two potential outcomes for unit i, one for each value of the endogenous regressor or treatment. Let W_i be the realized value of the endogenous regressor, equal to zero or one. We observe W_i and

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 1 \\ Y_i(0) & \text{if } W_i = 0. \end{cases}$$

Define two potential outcomes $W_i(0)$ and $W_i(1)$, representing the value of the endogenous regressor given the two values for the instrument Z_i. The actual or realized value of the endogenous variable is

$$W_i = W_i(Z_i) = \begin{cases} W_i(1) & \text{if } Z_i = 1 \\ W_i(0) & \text{if } Z_i = 0. \end{cases}$$

So we observe the triple Z_i, $W_i = W_i(Z_i)$ and $Y_i = Y_i(W_i(Z_i))$.

3. Local Average Treatment Effects

The key instrumental variables assumption is

Assumption 1 (Independence)

$$Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), W_i(0), W_i(1)).$$

It requires that the instrument is as good as randomly assigned, and that it does not directly affect the outcome. The assumption is formulated in a nonparametric way, without definitions of residuals that are tied to functional forms.

Assumptions (ctd)

Alternatively, we separate the assumption by postulating the existence of four potential outcomes, $Y_i(z, w)$, corresponding to the outcome that would be observed if the instrument was $Z_i = z$ and the treatment was $W_i = w$.

Assumption 2 (Random Assignment)

$$Z_i \perp\!\!\!\perp (Y_i(0,0), Y_i(0,1), Y_i(1,0), Y_i(1,1), W_i(0), W_i(1)).$$

and

Assumption 3 (Exclusion Restriction)

$$Y_i(z, w) = Y_i(z', w), \quad \text{for all } z, z', w.$$

The first of these two assumptions is implied by random assignment of Z_i, but the second is substantive, and randomization has no bearing on it.

Compliance Types

It is useful for our approach to think about the compliance behavior of the different units

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker	defier
	1	complier	always-taker

We cannot directly establish the type of a unit based on what we observe for them since we only see the pair (Z_i, W_i), not the pair $(W_i(0), W_i(1))$. Nevertheless, we can rule out some possibilities.

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker/defier
	1	always-taker/defier	complier/always-taker

Monotonicity

Assumption 4 (Monotonicity/No-Defiers)

$$W_i(1) \geq W_i(0).$$

This assumption makes sense in a lot of applications. It is implied directly by many (constant coefficient) latent index models of the type:

$$W_i(z) = 1\{\pi_0 + \pi_1 \cdot z + \varepsilon_i > 0\},$$

but it is much weaker than that.

Implications for Compliance types:

		Z_i = 0	Z_i = 1
W_i	0	complier/never-taker	never-taker
	1	always-taker	complier/always-taker

For individuals with $(Z_i = 0, W_i = 1)$ and for $(Z_i = 1, W_i = 0)$ we can now infer the compliance type.

Distribution of Compliance Types

Under random assignment and monotonicity we can estimate the distribution of compliance types:

$$\pi_a = \Pr(W_i(0) = W_i(1) = 1) = \mathbb{E}[W_i | Z_i = 0]$$

$$\pi_c = \Pr(W_i(0) = 0, W_i(1) = 1) = \mathbb{E}[W_i | Z_i = 1] - \mathbb{E}[W_i | Z_i = 0]$$

$$\pi_n = \Pr(W_i(0) = W_i(1) = 0) = 1 - \mathbb{E}[W_i | Z_i = 1]$$

Now consider average outcomes by instrument and treatment:

$$\mathbb{E}[Y_i|W_i=0,Z_i=0] =$$

$$\frac{\pi_c}{\pi_c+\pi_n}\cdot\mathbb{E}[Y_i(0)|\text{complier}]+\frac{\pi_n}{\pi_c+\pi_n}\cdot\mathbb{E}[Y_i(0)|\text{never}-\text{taker}],$$

$$\mathbb{E}[Y_i|W_i=0,Z_i=1]=\mathbb{E}[Y_i(0)|\text{never}-\text{taker}],$$

$$\mathbb{E}[Y_i|W_i=1,Z_i=0]=\mathbb{E}[Y_i(1)|\text{always}-\text{taker}],$$

$$\mathbb{E}[Y_i|W_i=1,Z_i=1] =$$

$$\frac{\pi_c}{\pi_c+\pi_a}\cdot\mathbb{E}[Y_i(1)|\text{complier}]+\frac{\pi_a}{\pi_c+\pi_a}\cdot\mathbb{E}[Y_i(1)|\text{always}-\text{taker}].$$

From this we can infer the average outcome for compliers,

$$\mathbb{E}[Y_i(0)|\text{complier}], \quad \text{and} \quad \mathbb{E}[Y_i(1)|\text{complier}],$$

Local Average Treatment Effect Hence the instrumental variables estimand, the ratio of these two reduced form estimands, is equal to the local average treatment effect

$$\beta^{\text{IV}}=\frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[W_i|Z_i=1]-\mathbb{E}[W_i|Z_i=0]}$$

$$=\mathbb{E}[Y_i(1)-Y_i(0)|\text{complier}].$$

4. Extrapolating to the Full Population

We can estimate

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right]$$

We can learn from these averages whether there is any evidence of heterogeneity in outcomes by compliance status, by comparing the pair of average outcomes of $Y_i(0)$;

$$\mathbb{E}\left[Y_i(0)|\text{never}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(0)|\text{complier}\right],$$

and the pair of average outcomes of $Y_i(1)$:

$$\mathbb{E}\left[Y_i(1)|\text{always}-\text{taker}\right], \quad \text{and} \quad \mathbb{E}\left[Y_i(1)|\text{complier}\right].$$

If compliers, never-takers and always-takers are found to be substantially different in levels, then it appears much less plausible that the average effect for compliers is indicative of average effects for other compliance types.

5. Application: Angrist (1990) effect of military service

The simple ols regression leads to:

$$\begin{array}{rcccc}\widehat{\log(\text{earnings})}_i & = & 5.4364 & - & 0.0205\cdot\widehat{\text{veteran}}_i\\ & & (0079) & & (0.0167)\end{array}$$

In Table we present population sizes of the four treatmen/instrument samples. For example, with a low lottery number 5,948 individuals do not, and 1,372 individuals do serve in the military.

		Z_i = 0	Z_i = 1
W_i	0	5,948	1,915
	1	1,372	865

Using these data we get the following proportions of the various compliance types, given in Table , under the non-defiers assumption. For example, the proportion of nevertakers is estimated as the conditional probability of $W_i = 0$ given $Z_i = 1$:

$$\Pr(\text{nevertaker}) = \frac{1915}{1915 + 865}.$$

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	never-taker (0.6888)	defier (0)
	1	complier (0.1237)	always-taker (0.1875)

Estimated Average Outcomes by Treatment and Instrument

		Z_i = 0	Z_i = 1
W_i	0	$\widehat{\mathbb{E}[Y]} = 5.4472$	$\widehat{\mathbb{E}[Y]} = 5.4028$
	1	$\widehat{\mathbb{E}[Y]} = 5.4076$,	$\widehat{\mathbb{E}[Y]} = 5.4289$

Not much variation by treatment status given instrument, but these comparisons are not causal under IV assumptions.

		$W_i(0)$ = 0	$W_i(0)$ = 1
$W_i(1)$	0	$\widehat{\mathbb{E}[Y_i(0)]} = 5.4028$	defier (NA)
	1	$\widehat{\mathbb{E}[Y_i(0)]} = 5.6948$, $\widehat{\mathbb{E}[Y_i(1)]} = 5.4612$	$\widehat{\mathbb{E}[Y_i(1)]} = 5.4076$

The local average treatment effect is -0.2336, a 23% drop in earnings as a result of serving in the military.

Simply doing IV or TSLS would give you the same numerical results:

$$\begin{array}{rcccc} \widehat{\log(\text{earnings})}_i & = & 5.4836 & - & 0.2336 \cdot \widehat{\text{veteran}}_i \\ & & (0.0289) & & (0.1266) \end{array}$$

It is interesting in this application to inspect the average outcome for different compliance groups. Average log earnings for never-takers are 5.40, lower by 29% than average earnings for compliers who do not serve in the military.

This suggests that never-takers are substantially different than compliers, and that the average effect of 23% for compliers need not be informative never-takers.

Note that
$\mathbb{E}[Y_i(0)|n, X_i] < \mathbb{E}[Y_i(0)|c, X_i]$,
but also $\mathbb{E}[Y_i(1)|c, X_i] > \mathbb{E}[Y_i(1)|a, X_i]$

Compliers earn more than nevertakers when not serving, and more than always-takers when serving. Does not fit standard gaussian selection model.

Part II: Regression Discontinuity Designs

1. Introduction
2. Basics
3. Graphical Analyses
4. Local Linear Regression
5. Choosing the Bandwidth
6. Variance Estimation
7. Specification Checks

1. Introduction

A Regression Discontinuity (RD) Design is a powerful and widely applicable identification strategy.

Often access to, or incentives for participation in, a service or program is assigned based on transparent rules with criteria based on clear cutoff values, rather than on discretion of administrators.

Comparisons of individuals that are similar but on different sides of the cutoff point can be credible estimates of causal effects for a specific subpopulation.

Good for internal validity, not much external validity.

Long history in Psychology literature (Thistlewaite and Campbell, 1960), early work by Goldberger (1972), recent resurgence in economics.

2. Basics

Two potential outcomes $Y_i(0)$ and $Y_i(1)$,
causal effect $Y_i(1) - Y_i(0)$,
binary treatment indicator W_i, covariate X_i,
and the observed outcome equal to:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases} \tag{1}$$

At $X_i = c$ incentives to participate change.

Two cases, **Sharp Regression Discontinuity**:

$$W_i = 1\{X_i \geq c\}. \tag{2}$$

and **Fuzzy Regression Discontinuity Design**:

$$\lim_{x \downarrow c} \Pr(W_i = 1 | X_i = x) \neq \lim_{x \uparrow c} \Pr(W_i = 1 | X_i = x), \tag{3}$$

Sharp Regression Discontinuity

Example (Lee, 2007)

What is effect of incumbency on election outcomes? (More specifically, what is the probability of a Democrat winning the next election given that the last election was won by a Democrat?)

Compare election outcomes in cases where previous election was very close.

SRD

Key assumption:
$\mathbb{E}[Y(0)|X = x]$ and $\mathbb{E}[Y(1)|X = x]$ are continuous in x.

Under this assumption,

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]. \quad (4)$$

The estimand is the difference of two regression functions at a point.

Extrapolation is unavoidable.

Fuzzy Regression Discontinuity

Examples (VanderKlaauw, 2002)

What is effect of financial aid offer on acceptance of college admission.

College admissions office puts applicants in a few categories based on numerical score.

Financial aid offer is highly correlated with category.

Compare individuals close to cutoff score.

FRD

What do we look at in the FRD case: ratio of discontinuities in regression function of outcome and treatment:

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i|X_i = x]}. \quad (5)$$

Interpretation of FRD (Hahn, Todd, VanderKlaauw)

Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neigborhood around c (which requires that the cutoff point is at least in principle manipulable)

$W_i(x)$ is non-increasing in x at $x = c$.

A complier is a unit such that

$$\lim_{x \downarrow X_i} W_i(x) = 0, \quad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

Then

$$\frac{\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[W_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i|X_i = x]}$$

$$= \mathbb{E}[Y_i(1) - Y_i(0)|\text{unit } i \text{ is a complier and } X_i = c].$$

External Validity

The estimatand has little external validity. It is at best valid for a population defined by the cutoff value c, and by the subpopulation that is affected at that value.

FRD versus Unconfoundedness

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i \mid X_i. \tag{6}$$

Under this assumption:

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c] = \mathbb{E}[Y_i|W_i = 1, X_i = c] - \mathbb{E}[Y_i|W_i = 0, X_i = c]$$

This approach assumes that differences between treated and control units with $X_i = c$ have a causal interpretation, without exploiting the discontinuity.

Unconfoundedness is fundamentally based on units being comparable if their covariates are similar. This is not an attractive assumption in the current setting where the probability of receiving the treatment is discontinuous in the covariate.

Even if unconfoundedness holds, under continuity of potential outcome regression functions FRD approach will be consistent for the average effect for compliers at $X_i = c$.

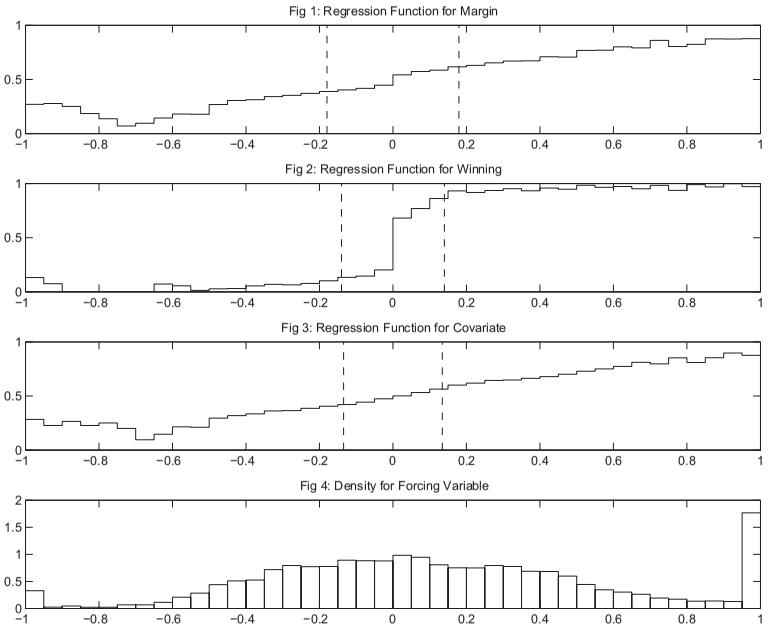
3. Graphical Analyses

A. Plot regression function $\mathbb{E}[Y_i|X_i = x]$

B. Plot regression functions $\mathbb{E}[Z_i|X_i = x]$ for covariates that do not enter the assignment rule Z_i

C. Plot density $f_X(x)$.

In all cases use estimators that do not smooth around the cutoff value. For example, for binwidth h define bins $[b_{k-1}, b_k]$, where $b_k = c - (K_0 - k + 1) \cdot h$, and average outcomes within bins.



4. Local Linear Regression

We are interested in the value of a regression function at the boundary of the support. Standard kernel regression

$$\widehat{\mu_l(c)} = \sum_{i|c-h<X_i<c}^{N} Y_i \Big/ \sum_{i|c-h<X_i<c}^{N} 1 \tag{7}$$

does not work well for that case (slower convergence rates)

Better rates are obtained by using local linear regression. First

$$\min_{\alpha_l,\beta_l} \sum_{i|c-h<X_i<c}^{N} (Y_i - \alpha_l - \beta_l \cdot (X_i - c))^2, \tag{8}$$

The value of lefthand limit $\mu_l(c)$ is then estimated as

$$\widehat{\mu_l(c)} = \widehat{\alpha}_l + \widehat{\beta}_l \cdot (c - c) = \widehat{\alpha}_l. \tag{9}$$

Similarly for righthand side. Not much gained by using a non-uniform kernel.

Alternatively one can estimate the average effect directly in a single regression,

$$Y_i = \alpha + \beta \cdot (X_i - c) + \tau \cdot W_i + \gamma \cdot (X_i - c) \cdot W_i + \varepsilon_i$$

thus solving

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} 1\{c - h \le X_i \le c + h\}$$

$$\times (Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i)^2,$$

which will numerically yield the same estimate of τ_{SRD}.

This interpretation extends easily to the inclusion of covariates.

Estimation for the FRD Case

Do local linear regression for both the outcome and the treatment indicator, on both sides,

$$\left(\widehat{\alpha}_{yl}, \widehat{\beta}_{yl}\right) = \arg \min_{\alpha_{yl},\beta_{yl}} \sum_{i:c-h\le X_i<c} \left(Y_i - \alpha_{yl} - \beta_{yl} \cdot (X_i - c)\right)^2,$$

$$\left(\widehat{\alpha}_{wl}, \widehat{\beta}_{wl}\right) = \arg \min_{\alpha_{wl},\beta_{wl}} \sum_{i:c-h\le X_i<c} (W_i - \alpha_{wl} - \beta_{wl} \cdot (X_i - c))^2,$$

and similarly $(\widehat{\alpha}_{yr}, \widehat{\beta}_{yr})$ and $(\widehat{\alpha}_{wr}, \widehat{\beta}_{wr})$. Then the FRD estimator is

$$\widehat{\tau}_{\mathsf{FRD}} = \frac{\widehat{\tau}_y}{\widehat{\tau}_w} = \frac{\widehat{\alpha}_{yr} - \widehat{\alpha}_{yl}}{\widehat{\alpha}_{wr} - \widehat{\alpha}_{wl}}.$$

Alternatively, define the vector of covariates

$$V_i = \begin{pmatrix} 1 \\ 1\{X_i < c\} \cdot (X_i - c) \\ 1\{X_i \geq c\} \cdot (X_i - c) \end{pmatrix}, \quad \text{and} \quad \delta = \begin{pmatrix} \alpha_{yl} \\ \beta_{yl} \\ \beta_{yr} \end{pmatrix}.$$

Then we can write

$$Y_i = \delta' V_i + \tau \cdot W_i + \varepsilon_i. \tag{10}$$

Then estimating τ based on the regression function (TSLS) by Two-Stage-Least-Squares methods, using

W_i as the endogenous regressor,
the indicator $1\{X_i \geq c\}$ as the excluded instrument
V_i as the set of exogenous variables

This is is numerically identical to $\widehat{\tau}_{\text{FRD}}$ before (because of uniform kernel)

Can add other covariates in straightfoward manner.

5. Choosing the Bandwidth (Imbens-Kalyanaraman)

We wish to take into account that (i) we are interested in the regression function at the boundary of the support, and (ii) that we are interested in the regression function at $x = c$.

IK focus on minimizing

$$\mathbb{E}\left[(\widehat{\mu}_l(c) - \widehat{\mu}_r(c) - (\mu_l(c) - \mu_r(c)))^2\right]$$

Both $\widehat{\mu}_l(c)$ and $\widehat{\mu}_r(c)$ are based on local linear estimators, with the same bandwidth h.

Optimal Bandwidth

$$h_{\text{opt}} = \left(\frac{C_2}{C_1}\right)^{1/5} \cdot \left(\frac{2 \cdot \sigma^2(c)/f_X(c)}{\left(\frac{\partial^2 m_r}{\partial x^2}(c) - \frac{\partial^2 m_l}{\partial x^2}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}.$$

where

$$C_1 = \frac{1}{4} \cdot \left(\frac{\nu_2^2 - \nu_1 \nu_3}{\nu_2 \nu_0 - \nu_1^2}\right)^2 \qquad C_2 = \frac{\nu_2^2 \pi_0 - 2\nu_1 \nu_2 \pi_1 + \nu_1^2 \pi_2}{(\nu_2 \nu_0 - \nu_1^2)^2}$$

$$\nu_j = \int_0^\infty u^j K(u) du, \quad \text{and} \quad \pi_j = \int_0^\infty u^j K^2(u) du.$$

If $K(u) = 1_{|u|<0.5}$, then $(C_2/C_1) = 5.40$

Bandwidth for FRD Design

1. Calculate optimal bandwidth separately for both regression functions and choose smallest.

2. Calculate optimal bandwidth only for outcome and use that for both regression functions.

Typically the regression function for the treatment indicator is flatter than the regression function for the outcome away from the discontinuity point (completely flat in the SRD case). So using same criterion would lead to larger bandwidth for estimation of regression function for treatment indicator. In practice it is easier to use the same bandwidth, and so to avoid bias, use the bandwidth from criterion for SRD design or smallest.

6. Variance Estimation

$$\sigma^2_{Yl} = \lim_{x \uparrow c} \mathsf{Var}(Y_i | X_i = x), \quad C_{YWl} = \lim_{x \uparrow c} \mathsf{Cov}(Y_i, W_i | X_i = x),$$

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl}\right), \quad V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl}\right)$$

The asymptotic covar of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ and $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$C_{\tau_y, \tau_w} = \frac{4}{f_X(c)} \cdot (C_{YWr} + C_{YWl}).$$

Finally, the asymptotic distribution has the form

$$\sqrt{Nh} \cdot (\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\tau_w^2} \cdot V_{\tau_y} + \frac{\tau_y^2}{\tau_w^4} \cdot V_{\tau_w} - 2 \cdot \frac{\tau_y}{\tau_w^3} \cdot C_{\tau_y, \tau_w}\right).$$

This asymptotic distribution is a special case of that in HTV, using the rectangular kernel, and with $h = N^{-\delta}$, for $1/5 < \delta < 2/5$ (so that the asymptotic bias can be ignored).

Can use plug in estimators for components of variance.

TSLS Variance for FRD Design

The second estimator for the asymptotic variance of $\hat{\tau}$ exploits the interpretation of the $\hat{\tau}$ as a TSLS estimator.

The variance estimator is equal to the robust variance for TSLS based on the subsample of observations with $c - h \leq X_i \leq c + h$, using the indicator $1\{X_i \geq c\}$ as the excluded instrument, the treatment W_i as the endogenous regressor and the V_i as the exogenous covariates.

7. Concerns about Validity

Two main conceptual concerns in the application of RD designs, sharp or fuzzy.

Other Changes

Possibility of other changes at the same cutoff value of the covariate. Such changes may affect the outcome, and these effects may be attributed erroneously to the treatment of interest.

Manipulation of Forcing Variable

The second concern is that of manipulation of the covariate value.

Specification Checks

A. Discontinuities in Average Covariates

B. A Discontinuity in the Distribution of the Forcing Variable

C. Discontinuities in Average Outcomes at Other Values

D. Sensitivity to Bandwidth Choice

7.A Discontinuities in Average Covariates

Test the null hypothesis of a zero average effect on pseudo outcomes known not to be affected by the treatment.

Such variables includes covariates that are by definition not affected by the treatment. Such tests are familiar from settings with identification based on unconfoundedness assumptions.

Although not required for the validity of the design, in most cases, the reason for the discontinuity in the probability of the treatment does not suggest a discontinuity in the average value of covariates. If we find such a discontinuity, it typically casts doubt on the assumptions underlying the RD design.

7.B A Discontinuity in the Distribution of the Forcing Variable

McCrary (2007) suggests testing the null hypothesis of continuity of the density of the covariate that underlies the assignment at the discontinuity point, against the alternative of a jump in the density function at that point.

Again, in principle, the design does not require continuity of the density of X at c, but a discontinuity is suggestive of violations of the no-manipulation assumption.

If in fact individuals partly manage to manipulate the value of X in order to be on one side of the boundary rather than the other, one might expect to see a discontinuity in this density at the discontinuity point.

Illustration Based on David Lee Election Data

Forcing variable is difference in dem vs rep vote share in last election.

Outcomes are dem vote share in next election, and indicator for democrats winning the next election.

Covariate for testing is dem vote share in prior election

6558 congressional elections.

Edge kernel: $K(u) = (1-u) \cdot 1_{0 \leq u \leq 1}$

Outcome	IK Bandwidth	Estimate	(s.e.)
Dem Win Next Elect	0.35	0.082	(0.007)
Demt Margin Next Election	0.20	0.411	(0.033)
Dem Margin Prev Election	0.28	0.005	(0.009)

AEA Lectures **Chicago, IL, January 2012**

Lecture 5, Monday, Jan 9th,pm-pm

Discrete Choice Models

1. Introduction

In this lecture we discuss multinomial discrete choice models. The modern literature on these models goes back to the work by Daniel McFadden in the seventies and eighties, (McFadden, 1973, 1981, 1982, 1984). In the nineties these models received much attention in the Industrial Organization literature, starting with Berry (1994), Berry, Levinsohn, Pakes (1995, BLP), and Goldberg (1995). In the IO literature the applications focused on demand for differentiated products, in settings with relatively large numbers of products, some of them close substitutes. In these settings a key feature of the conditional logit model, namely the Independence of Irrelevant Alternatives (IIA), was viewed as particularly unattractive. Three approaches have been used to deal with this. Goldberg (1995) used nested logit models to avoid the IIA property. McCulloch and Rossi (1994), and McCulloch, Polson and Rossi (2000) studied multinomial probit models with relatively unrestricted covariance matrices for the unobserved components. BLP, McFadden and Train (2000) and Berry, Levinsohn and Pakes (2004) uses random effects or mixed logit models, in BLP in combination with unobserved choice characteristics and using methods that allow for estimation using only aggregate choice data. The BLP approach has been very influential in the subsequent empirical IO literature.

Here we discuss these models. We argue that the random effects approach to avoid IIA is indeed very attractive, both substantively and computationally, compared to the nested logit or unrestricted multinomial probit models. In addition to the use of random effects to avoid the IIA property, the inclusion in the BLP methodology of unobserved choice characteristics, and the ability to estimate the models with market share rather than individual level data makes their methods very flexible and widely applicable. We discuss extensions to the BLP set up allowing multiple unobserved choice characteristics, and the richness required for these

models to rationalize general choice data based on utility maximization. We also discuss the potential benefits of using Bayesian methods.

2. Multinomial and Conditional Logit Models

First we briefly review the multinomial and conditional logit models.

2.1 Multinomial Logit Models

We focus on models for discrete choice with more than two choices. We assume that the outcome of interest, the choice Y_i takes on non-negative, un-ordered integer values between zero and J; $Y_i \in \{0, 1, \ldots, J\}$. Unlike the ordered case there is no particular meaning to the ordering. Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan), and many others.

We wish to model the distribution of Y in terms of covariates. In some cases we will distinguish between covariates Z_i that vary by units (individuals or firms), and covariates that vary by choice (and possibly by individual), X_{ij}. Examples of the first type include individual characteristics such as age or education. An example of the second type is the cost associated with the choice, for example the cost of commuting by bus/train/car, or the price of a product, or the speed of a computer chip. This distinction is important from the substantive side of the problem. McFadden developed the interpretation of these models through utility maximizing choice behavior. In that case we may be willing to put restrictions on the way covariates affect utilities: characteristics of a particular choice should affect the utility of that choice, but not the utilities of other choices.

The strategy is to develop a model for the conditional probability of choice j given the covariates. Suppose we only have individual-specific covariates, and the model is $\Pr(Y_i = j|Z_i = z) = P_j(z;\theta)$. Then the log likelihood function is

$$L(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{J} \mathbf{1}_{Y_i=j} \cdot \ln P_j(Z_i;\theta).$$

A natural extension of the binary logit model is to model the response probability as

$$\Pr(Y_i = j|Z_i = z) = \frac{\exp(z'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for choices $j = 1, \ldots, J$ and

$$\Pr(Y_i = 0|Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for the first choice. The γ_l here are choice-specific parameters. This multinomial logit model leads to a very well-behaved likelihood function, and it is easy to estimate using standard optimization techniques. Interestingly, it can be viewed as a special case of the following conditional logit.

2.2 Conditional Logit Models

Suppose all covariates vary by choice (and possibly also by individual, but that is not essential here). Then McFadden proposed the conditional logit model:

$$\Pr(Y_i = j|X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)},$$

for $j = 0, \ldots, J$. Now the parameter vector β is common to all choices, and the covariates are choice-specific.

The multinomial logit model can be viewed as a special case of the conditional logit model. Suppose we have a vector of individual characteristics Z_i of dimension K, and J vectors of coefficients γ_j, each of dimension K. Then define for choice j, $j = 1, \ldots, J$, the vector of covariates X_{ij} as the vector of dimension $K \times J$, with all elements equal to zero other than the elements $K \times (j-1) + 1$ to $K \times j$ which are equal to Z_i:

$$X_{i1} = \begin{pmatrix} Z_i \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{ij} = \begin{pmatrix} 0 \\ \vdots \\ Z_i \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots \quad X_{iJ} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ Z_i \end{pmatrix}, \text{ and } X_{i0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and define the common parameter vector β, of dimension $K \cdot J$, as

$$\beta = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_J \end{pmatrix}.$$

Then

$$\Pr(Y_i = j|Z_i) = \frac{\exp(Z_i'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = j|X_{i0}, \ldots, X_{iJ}),$$

for $j = 1, \ldots, J$, and

$$\Pr(Y_i = 0|Z_i) = \frac{1}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)} = \frac{\exp(X_{i0}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = 0|X_{i0}, \ldots, X_{iJ}).$$

2.3 Link with Utility Maximization

McFadden motivates the conditional logit model by extending the single latent index model to multiple choices. Suppose that the utility, for individual i, associated with choice j, is

$$U_{ij} = X_{ij}'\beta + \varepsilon_{ij}. \tag{1}$$

Furthermore, let individual i choose option j (so that $Y_i = j$) if choice j provides the highest level of utility, or

$$Y_i = j \quad \text{if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

(ties have probability zero because of the continuity of the distribution for ε).

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions. Then the choice Y_i follows the conditional logit model. The type I extreme value distribution has cumulative distribution function

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad \text{and pdf} \quad f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

This distribution has a unique mode at zero, a mean equal to 0.58, and a a second moment of 1.99 and a variance of 1.65. See Figure 1 for the probability density function and the comparison with the normal density. Note the assymmetry of the distribution.

Given the extreme value distribution the probability of choice 0 is

$$\begin{aligned}
\Pr(Y_i = 0|X_{i0},\ldots,X_{iJ}) &= \Pr(U_{i0} > U_{i1},\ldots,U_{i0} > U_{iJ}) \\
&= \Pr(\varepsilon_{i0} + X_{i0}'\beta - X_{i1}'\beta > \varepsilon_{i1},\ldots,\varepsilon_{i0} + X_{i0}'\beta - X_{iJ}'\beta > \varepsilon_{iJ}) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_{i0}+X_{i0}'\beta-X_{i1}'\beta}\cdots\int_{-\infty}^{\varepsilon_{i0}+X_{iJ}'\beta-X_{iJ}'\beta} f(\varepsilon_{i0})\ldots f(\varepsilon_{iJ})d\varepsilon_{iJ}\ldots,d\varepsilon_{i0} \\
&= \int_{-\infty}^{\infty}\exp(-\varepsilon_{0i})\exp(-\exp(-\varepsilon_{0i})\cdot\exp(-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{i1}'\beta))\ldots \\
&\qquad\qquad \times\exp(-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{iJ}'\beta))d\varepsilon_{i0} \\
&= \int_{-\infty}^{\infty}\exp(-\varepsilon_{0i})\exp\Big[-\exp(-\varepsilon_{0i})-\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{i1}'\beta))\ldots \\
&\qquad\qquad -\exp(-\varepsilon_{i0}-X_{i0}'\beta+X_{iJ}'\beta)\Big]d\varepsilon_{i0} \\
&= \frac{\exp(X_{i0}'\beta)}{\sum_{j=0}^{J}\exp(X_{j0}'\beta)}.
\end{aligned}$$

To see the different steps in this derivation note that

$$\int_{-\infty}^{c}\exp(-\epsilon)\cdot\exp(-\exp(-\epsilon))d\epsilon = F(c) = \exp(-\exp(-c)),$$

for the extreme value distribution. Also,

$$\int_{-\infty}^{\infty}\exp(-\epsilon)\cdot\exp(-\exp(-\epsilon-c))d\epsilon$$

$$= \int_{-\infty}^{\infty} \exp(-\eta + c) \cdot \exp(-\exp(-\eta))d\eta$$

$$= \exp(c) \cdot \int_{-\infty}^{\infty} \exp(-\eta) \cdot \exp(-\exp(-\eta))d\eta = \exp(c),$$

by change of variables, which we apply with

$$c = -\ln\left(1 + \exp(X_{i1}'\beta - X_{i0}'\beta) + \ldots + \exp(X_{iJ}'\beta - X_{i0}'\beta)\right).$$

3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA). Consider the conditional probability of choosing j given that you choose either j or l:

$$\Pr(Y_i = j | Y_i \in \{j, l\}) = \frac{\Pr(Y_i = j)}{\Pr(Y_i = j) + \Pr(Y_i = l)} = \frac{\exp(X_{ij}'\beta)}{\exp(X_{ij}'\beta) + \exp(X_{il}'\beta)}.$$

This probability does not depend on the characteristics X_{im} of alternatives m other than j and l. This is sometimes unattractive. The traditional example is McFadden's famous blue bus/red bus example. Suppose there are initially three choices: commuting by car, by red bus or by blue bus. It would seem reasonable be to assume that people have a preference over cars versus buses, but are indifferent between red versus blue buses. One could capture this by assuming that

$$U_{i,\text{redbus}} = U_{i,\text{bluebus}},$$

with the choice between the blue and red bus being random. So, to be explicit, suppose that $X_{i,\text{bluebus}} = X_{i,\text{redbus}} = X_{i,\text{bus}}$. Then suppose that the probability of commuting by bus is

$$\Pr(Y_i = \text{bus}) = \Pr(Y_i = \text{redbus or bluebus}) = \frac{\exp(X_{i,\text{bus}}'\beta)}{\exp(X_{i,\text{bus}}'\beta) + \exp(X_{i,\text{car}}'\beta)},$$

and the probability of choosing a red bus or blue bus, conditional on choosing a bus, is

$$\Pr(Y_i = \text{redbus} | Y_i = \text{bus}) = \frac{1}{2}.$$

That would imply that the conditional probability of commuting by car, given that one commutes by car or red bus, would differ from the same conditional probability if there is no blue bus. Presumably taking away the blue bus choice would lead all the current blue bus users to shift to the red bus, and not to cars.

The conditional logit model does not allow for this type of substitution pattern. Another way of stating the problems with the conditional logit model is to say that it generates unrealistic substitution patterns. Let us make that argument more specific. Suppose that individuals have the choice out of three Berkeley restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose the two characteristics of the restaurants are price with $P_C = 95$, $P_L = 80$, and $P_B = 5$, and quality, with $Q_C = 10$, $Q_L = 9$, and $Q_B = 2$. Suppose that market shares for the three restaurants are $S_C = 0.10$, $S_L = 0.25$, and $S_B = 0.65$. These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

$$U_{ij} = -0.2 \cdot P_j + 2 \cdot Q_j + \epsilon_{ij},$$

with independent extreme value ϵ_{ij}, and individuals go to the restaurant with the highest utility. Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business). In that case the prediction of the conditional logit model is that the market shares for Chez Panisse and the Bongo Burger go to $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$. That seems implausible. The people who were planning to go to Lalime's would appear to be more likely to go to Chez Panisse if Lalime's is closed than to go to the Bongo Burger, and so one would expect $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$. The model on the other hand predicts that most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

Recall the latent utility set up with the utility for individual i and choice j equal to

$$U_{ij} = X'_{ij}\beta + \epsilon_{ij}. \tag{2}$$

In the conditional logit model we assume independent ϵ_{ij} with extreme value distributions. This is essentially what creates the IIA property. (This is not completely correct, because other distributions for the unobserved, say with normal errors, we would not get IIA exactly, but something pretty close to it.) The solution is to allow in some fashion for correlation between the unobserved components in the latent utility representation. In particular, with a choice set that contains multiple versions of essentially the same choice (like the red bus or the blue bus), we should allow the latent utilities for these choices to be identical, or at least very close. In order to achieve this the unobserved components of the latent utilities would have to be highly correlated for those choices. This can be done in a number of ways.

4. Models without Independence of Irrelevant Alternatives

Here we discuss three ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the unobserved components of the latent utility. All of these originate in some form or another in McFadden's work (e.g., McFadden, 1981, 1982, 1984). The first is the nested logit model where the researcher groups together sets of choices. In the simple version with a single layer of nests this allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation between the unobserved components of choices in different nests. Second, the unrestricted multinomial probit model with no restrictions on the covariance between unobserved components, beyond normalizations. Third, the mixed or random coefficients logit where the marginal utilities associated with choice characteristics are allowed to vary between individuals. This generates positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

4.1 Nested Logit

One way to induce correlation between the choices is through nesting them. Suppose the

set of choices $\{0, 1, \ldots, J\}$ can be partitioned into S sets $B_1, \ldots, B_S$, so that the full set of choices can be written as

$$\{0, 1, \ldots, J\} = \cup_{s=1}^{S} B_s.$$

Let Z_s be set-specific characteristics. (It may be that the set of set specific variables is empty, or just a vector of indicators, with Z_s an S-vector of zeros with a one for the sth element.) Now let the conditional probability of choice j given that your choice is in the set B_s, or $Y_i \in B_s$ be equal to

$$\Pr(Y_i = j | X_i, Y_i \in B_s) = \frac{\exp(\rho_s^{-1} X_{ij}'\beta)}{\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}'\beta)},$$

for $j \in B_s$, and zero otherwise. In addition suppose the marginal probability of a chocie in the set B_s is

$$\Pr(Y_i \in B_s | X_i) = \frac{\exp(Z_s'\alpha) \left(\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}'\beta)\right)^{\rho_s}}{\sum_{t=1}^{S} \exp(Z_t'\alpha) \left(\sum_{l \in B_t} \exp(\rho_t^{-1} X_{il}'\beta)\right)^{\rho_s}}.$$

If we fix $\rho_s = 1$ for all s, then

$$\Pr(Y_i = j | X_i) = \frac{\exp(X_{ij}'\beta + Z_s'\alpha)}{\sum_{t=1}^{S} \sum_{l \in B_t} \exp(X_{il}'\beta + Z_t\alpha)},$$

and we are back in the conditional logit model.

In general this model corresponds to individuals choosing the option with the highest utility, where the utility of choice j in set B_s for individual i is

$$U_{ij} = X_{ij}'\beta + Z_s'\alpha + \epsilon_{ij},$$

where the joint distribution function of the ϵ_{ij} is

$$F(\epsilon_{i0}, \ldots, \epsilon_{iJ}) = \exp\left(-\sum_{s=1}^{S} \left(\sum_{j \in B_s} \exp\left(-\rho_s^{-1}\epsilon_{ij}\right)\right)^{\rho_s}\right).$$

Within the sets the correlation coefficient for the ϵ_{ij} is approximately equal to $1-\rho$. Between the sets the ϵ_{ij} are independent.

The nested logit model could capture the blue bus/red bus example by having two nests, the first $B_1 = \{\text{redbus}, \text{bluebus}\}$, and the second one $B_2 = \{\text{car}\}$.

How do you estimate these models? One approach is to construct the log likelihood and directly maximize it. That is complicated, especially since the log likelihood function is not concave, but it is not impossible. An easier alternative is to directly use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Hence we can directly estimate β/ρ_s using the concavity of the conditional logit model. Denote these estimates of β/ρ_s by $\widehat{\beta/\rho_s}$. Then the probability of a particular set B_s can be used to estimate ρ_s and α through

$$\Pr(Y_i \in B_s|X_i) = \frac{\exp(Z_s'\alpha)\left(\sum_{l\in B_s}\exp(X_{il}'\widehat{\beta/\rho_s})\right)^{\rho_s}}{\sum_{t=1}^S \exp(Z_t'\alpha)\left(\sum_{l\in B_t}\exp(X_{il}'\widehat{\beta/\rho_t})\right)^{\rho_s}} = \frac{\exp(Z_s'\alpha+\rho_s\hat{W}_s)}{\sum_{t=1}^S\exp(Z_t'\alpha+\rho_t\hat{W}_t)},$$

where

$$\hat{W}_s = \ln\left(\sum_{l\in B_s}\exp(X_{il}'\widehat{\beta/\rho_s})\right),$$

known as the "inclusive values". Hence we have another conditional logit model back that is easily estimable. These two-step estimators are not efficient. The variance/covariance matrix is provided in McFadden (1981).

These models can be extended to many layers of nests. See for an impressive example of a complex model with four layers of multiple nests Goldberg (1995). Figure 2 shows the nests in the Goldberg application. The key concern with the nested logit models is that results may be sensitive to the specification of the nest structure. The researcher chooses the choices that are potentially close, with the data being used to estimate the amount of correlation. In contrast, in the random effects models, choices can only be close if they are close in terms of observed choice characteristics, with the data being used to estimate the

relative importance of the various choice characteristics. In that sense the nested logit model can be more flexible, allowing the researcher to group together choices that are far apart in terms of observed choice characteristics, but it is more demanding in requiring the researcher to make these decisions *a priori*.

4.2 Multinomial Probit

A second possibility is to directly free up the covariance matrix of the error terms. This is more natural to do in the multinomial probit case. See McCulloch and Rossi (1994) McCulloch, Polson, and Rossi (2000) for general discussion.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i0}'\beta + \epsilon_{i0} \\ X_{i1}'\beta + \epsilon_{i1} \\ \vdots \\ X_{iJ}'\beta + \epsilon_{iJ} \end{pmatrix},$$

with

$$\epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} \Big| X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1) \times (J+1)$ covariance matrix Ω. We do need some normalizations on Ω beyond symmetry. Recall that in the binary choice case (which corresponds to $J = 1$) there were no free parameters in the distribution of ϵ, which implies three restrictions on the symmetric matrix Ω.

In principle we can derive the probability for each choice given the covariates, construct the likelihood function based on that, and maximize it using an optimization algorithm like Davidon-Fletcher-Powell (Gill, Murray, and Wright, 1981) or something similar. In practice this is very difficult with $J \geq 3$. Evaluating the probabilities involves calculating a third order integral involving normal densities. This is difficult to to using standard integration methods. There are two alternatives.

There is a substantial literature on simulation methods for computing estimates in these models. See for an early example Manski and Lerman (1981), general studies McFadden (1989), and Pakes and Pollard (1989), and Hajivassiliou and Ruud (1994) for a review. Geweke, Keane, and Runkle (1994) and Hajivasilliou and McFadden (1990) proposed a way of calculating the probabilities in the multinomial probit models that allowed researchers to deal with substantially larger choice sets. A simple attempt to estimate the probabilities would be to draw the ϵ_i from a multivariate normal distribution and calculate the probability of choice j as the number of times choice j corresponded to the highest utility. This does not work well in cases with many (more than four) choices. The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws sequentially and combines the draws with the calculation of univariate normal integrals so that the resulting probabilities are smooth in the parameters.

From a Bayesian perspective drawing from the posterior distribution of β and Ω is straightforward. The key is setting up the vector of unobserved random variables as

$$\theta = (\beta, \Omega, U_{i0}, \ldots, U_{iJ}) ,$$

and defining the most convenient partition of this vector. Suppose we know the latent utilities U_i for all individuals. Then the normality makes this a standard linear model problem, and we can sample sequentially from $\beta|\Omega$ and $\Omega|\beta$ given the appropriate conjugate prior distributions (normal for β and inverse Wishart for Ω). Given the parameters drawing from the unobserved utilities can be done sequentially: for each unobserved utility given the others we would have to draw from a truncated normal distribution, which is straightforward. See McCulloch, Polson, and Rossi (2000) for details.

The attraction of this approach is that there are no restrictions on which choices are close. In contrast, in the nested logit approach the researcher specifies which choices are potentially close, and in the random effects approach only choices that are close in terms of observed choice characteristics can be close. The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of chocies this frees up a

large number of parameters (all elements in the $(J+1) \times (J+1)$ dimensional covariance matrix of latent utilities, minus some that are fixed by normalizations.) Estimating all these covariance parameters precisely, based on only first choice data (as opposed to data where we know for each individual additional orderings, e.g., first and second choices), is difficult with the sample sizes typically available.

4.3 Random Coefficient (Mixed) Logit (or Probit)

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients. This is a very natural idea. Why do we fundamentally think that if Lalime's price goes up, the individuals who were planning to go Lalime's go to Chez Panisse instead, rather than to the Bongo Burger? The reason is that we think individuals who have a taste for Lalime's are likely to have a taste for close substitute in terms of observable characteristics, Chez Panisse as well, rather than for the Bongo Burger.

We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X_{ij}'\beta_i + \epsilon_{ij},$$

where the ϵ_{ij} are again independent of everything else, and of each other, either extreme value, or normal. We can also write this as

$$U_{ij} = X_{ij}'\overline{\beta} + \nu_{ij},$$

where

$$\nu_{ij} = \epsilon_{ij} + X_{ij} \cdot (\beta_i - \overline{\beta}),$$

which is no longer independent across choices. The key ingredient is the vector of individual specific taste parameters β_i. See for a general discussion of such models and their properties in approximating general choice patterns McFadden and Train (2000). One possibility is to

assume the existence of a finite number of types of individuals, similar to the mixture models used by Heckman and Singer (1984) in duration settings:

$$\beta_i \in \{b_0, b_1, \ldots, b_K\},$$

with

$$\Pr(\beta_i = b_k|Z_i) = p_k, \qquad \text{or} \quad \Pr(\beta_i = b_k|Z_i) = \frac{\exp(Z_i'\gamma_k)}{1 + \sum_{l=1}^{K} \exp(Z_i'\gamma_l)}.$$

Here the taste parameters take on a finite number of values, and we have a finite mixture. We can use either Gibbs sampling with the indicator of which mixture an observations belongs to as an unobserved random variable, or use the EM algorithm (Dempster, Laird, and Rubin, 1977).

Alternatively we could specify

$$\beta_i|Z_i \sim \mathcal{N}(Z_i'\gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters. Just evaluating the likelihood function would be very difficult in this setting if there is a large number of choices. This would involve integrating out the random coefficients which could be very computationally intensive. See McFadden and Train (2000). Using Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

5. Berry-Levinsohn-Pakes

Here we consider again random effects logit models. BLP extended these models to allow for unobserved product characteristics, endogeneity of choice characteristics, and developed methods that allowed for consistent estimation without individual level choice data. Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products, often in settings with a large number of products. See Nevo (2000) and Ackerberg, Benkard, Berry, and Pakes (2005) for reviews and references.

Compared to the earlier examples we have looked at there is an emphasis in this study, and those that followed it, on the large number of goods and the potential endogeneity of some of the product characteristics. (Typically one of the regressors is the price of the good.) In addition the procedure only requires market level data. We do not need individual level purchase data, just market shares and estimates of the distribution of individual characteristics by market. In practice we need a fair amount of variation in these things to estimate the parameters well, but in principle this is less demanding in terms of data required. On the other hand, we do need data by market, where before we just needed individual purchases in a single market (although to identify price effects we would need variation in prices by individuals in that case).

The data have three dimensions: products, indexed by $j = 0, \ldots, J$, markets, $t = 1, \ldots, T$, and individuals, $i = 1, \ldots, N_t$. We only observe one purchase per individual. The large sample approximations are based on large N and T, and fixed J.

Let us go back to the random coefficients model, now with each utility indexed by individual, product and market:

$$U_{ijt} = \beta_i' X_{jt} + \zeta_{jt} + \epsilon_{ijt}.$$

The ζ_{jt} is a unobserved product characteristic. This component is allowed to vary by market and product. It can include product and market dummies (for example, we can have $\zeta_{jt} = \zeta_j + \zeta_t$). Unlike the observed product characteristics this unobserved characteristic does not have a individual-specific coefficient. The inclusion of this component allows the model to rationalize any pattern of market shares. The observed product characteristics may include endogenous characteristics like the price.

The ϵ_{ijt} unobserved components have extreme value distributions, independent across all individuals i, products j, and markets t.

The random coefficients β_i, with dimension equal to that of the observable characteristics X_{jt}, say K, are assumed to be related to individual observable characteristics. We postulate

the following linear form:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i,$$

with

$$\eta_i | Z_i \sim \mathcal{N}(0, \Sigma).$$

So if the dimension of Z_i is $L \times 1$, then Γ is a $L \times K$ matrix. The Z_i are normalized to have mean zero, so that the β's are the average marginal utilities. The normality assumption is not necessary, and unlikely to be important. Other distributional assumptions can be substituted.

BLP developed an approach to estimate models of this type that does not require individual level data. Instead it exploits aggregate (market level) data in combination with estimates of the distribution of Z_i. Specifically the data consist of estimated shares $\hat{s}_{tj}$ for each choice j in each market t, combined with observations from the marginal distribution of individual characteristics (the Z_i's) for each market, often from representative data sets such as the CPS.

First write the latent utilities as

$$U_{ijt} = \delta_{jt} + \nu_{ijt} + \epsilon_{ijt},$$

where

$$\delta_{jt} = \beta' X_{jt} + \zeta_{jt}, \qquad \text{and} \quad \nu_{ijt} = (Z_i'\Gamma + \eta_i)' X_{jt}.$$

Now consider for fixed Γ and Σ and δ_{jt} the implied market share for product j in market t, s_{jt}. This can be calculated analytically in simple cases. For example with $\Gamma_{jt} = 0$ and $\Sigma = 0$, the market share is a very simple function of the δ_{jt}:

$$s_{jt}(\delta_{jt}, \Gamma = 0, \Sigma = 0) = \frac{\exp(\delta_{jt})}{\sum_{l=0}^{J} \exp(\delta_{lt})}.$$

More generally, this is a more complex relationship. We can always calculate the implied market share by simulation: draw from the distribution of Z_i in market t, draw from the distribution of η_i, and calculate the implied purchase probability (or even simulate the implied purchase by also drawing from the distribution of ϵ_{ijt}). Do that repeatedly and you will be able to calculate the market share for this product/market. Call the vector function obtained by stacking these functions for all products and markets $s(\delta, \Gamma, \Sigma)$.

Next, fix only Γ and Σ. For each value of δ_{jt} we can find the implied market share. Now find the vector of δ_{jt} such that the implied market shares are equal to the observed market shares $\hat{s}_{jt}$ for all j, t. BLP suggest using the following algorithm. Given a starting value for δ^0_{jt}, use the updating formula:

$$\delta^{k+1}_{jt} = \delta^k_{jt} + \ln s_{jt} - \ln s_{jt}(\delta^k, \Gamma, \Sigma).$$

BLP show this is a contraction mapping, and so it defines a function $\delta(s, \Gamma, \Sigma)$ expressing the δ as a function of observed market shares, and parameters Γ and Σ. In order to implement this, one needs to approximate the implied market shares accurately for each iteration in the contraction mapping, and then you will need to do this repeatedly to get the contraction mapping to converge.

Note that does require that each market share is accurately estimated. If all we have is an estimated market share, then even if this is unbiased, the procedures will not necessarily work. In that case the log of the estimated share is not unbiased for the log of the true share. In practice the precision of the estimated market share is so much higher than that of the other parameters that this is unlikely to matter.

Given this function $\delta(s, \Gamma, \Sigma)$ define the residuals

$$\omega_{jt} = \delta_{jt}(s, \Gamma, \Sigma) - \beta' X_{jt}.$$

At the true values of the parameters and the true market shares this is equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM or instrumental variable methods. We assume that the unobserved product characteristics are uncorrelated with observed product characteristics (other than typically price). This is not sufficient since the observed product characteristics enter directly into the model. We need more instruments, and typically use things like characteristics of other products by the same firm, or average characteristics by competing products. The general GMM machinery will also give us the standard errors for this procedure. This is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

It is instructive to see what this approach does if we in fact have, and know we have, a conditional logit model with fixed coefficients. In that case $\Gamma = 0$, and $\Sigma = 0$. Then we can invert the market share equation to get the market specific unobserved choice-characteristics

$$\delta_{jt} = \ln s_{jt} - \ln s_{0t},$$

where we set $\delta_{0t} = 0$. (this is typically the outside good, whose average utility is normalized to zero). The residual is

$$\zeta_{jt} = \delta_{jt} - \beta' X_{jt} = \ln s_{jt} - \ln s_{0t} - \beta' X_{jt}.$$

With a set of instruments W_{jt}, we run the regression

$$\ln s_{jt} - \ln s_{0t} = \beta' X_{jt} + \epsilon_{jt},$$

using W_{jt} as instrument for X_{jt}, using as the observational unit the market share for product j in market t.

So here the technique is very transparent. It amounts to transforming the market shares to something linear in the coefficients so we can use two-stage-least-squares. More generally the transformation is going to be much more difficult with the random coefficients implying that there is no analytic solution. Computationally these things can get very complicated. Note however that we can estimate these models now without having individual level data,

and that at the same time we can get a fairly flexible model for the substitution patterns. At the same time you would expect to need a lot of structure to get the parameters precisely estimated just as in the other models. Of course if you compare the current model to the nested logit model you can impose such structure by imposing restrictions on the covariance matrix.

Comparisons of the models are difficult. Obviously if the structure imposed is correct it helps, but we typically do not know what the truth is, so we cannot conclude which one is better on the basis of the data typically available.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach allows for a single unobserved choice characteristic. This is essential for their estimation strategy that requires only market share data, and exploits a one-to-one relationship between market-specific unobserved product characteristics and market shares given other parameters and covariates. With individual level data one may be able to, and wish to allow for, multiple unobserved product characteristics. Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007), among others, study such models, in all cases with the unobserved choice characteristics constant across markets. Athey and Imbens model the latent utility for individual i in market t for choice j as

$$U_{ijt} = X_{it}'\beta_i + \zeta_j'\gamma_i + \epsilon_{ijt},$$

with the individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega).$$

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult. Athey and Imbens show that Bayesian methods, and in particular markov-chain-monte-carlo methods are effective tools for conducting inference in these settings.

7. Hedonic Models and the Motivation for a Choice and Individual Specific Error Term

Recently researchers have reconsidered using pure characteristics models for discrete choices, that is models with no idiosyncratic error ϵ_{ij}, instead relying solely on the presence of a few unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices. Such an error term is the only source of stochastic variation in the original multinomial choice models with only observed choice and individual characteristics, but in models with unobserved choice and individual characteristics their presence needs more motivation. Athey and Imbens (2007) discuss two arguments for including the additive error term.

First, the pure characteristics model can be extremely sensitive to measurement error, because it can predict zero market shares for some products. Consider a case where choices are generated by a pure characteristics model that implies that a particular choice j has zero market share. Now suppose that there is a single unit i for whom we observe, due to measurement error, the choice $Y_i = j$. Irrespective of the number of correctly measured observations available that were generated by the pure characteristics model, the estimates of the latent utility function will not be close to the true values corresponding to the pure characteristics model due to the single mismeasured observation. Such extreme sensitivity puts a lot of emphasis on the correct specification of the model and the absence of measurement error, and is undesirable in most settings.

Thus, one might wish to generalize the model to be robust against small amounts of measurement error of this type. One possibility is to define the optimal choice Y_i^* as the choice that maximizes the utility and assume that the observed choice Y_i is equal to the optimal choice Y_i^* with probability $1-\delta$, and with probability $\delta/(J-1)$ any of the other choices is observed:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J) = \begin{cases} 1-\delta & \text{if } Y_i^* = y, \\ \delta/(J-1) & \text{if } Y_i^* \neq y. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without having the disad-

vantages of extreme sensitivity to mismeasured choices that the pure characteristics model has. If the true choices are generated by the pure characteristics model the presence of a single mismeasured observation will not prevent the researcher from estimating the true utility function. However, this specific generalization of the pure characteristics model has an unattractive feature: if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely. One might expect that choices with utilities closer to the optimal one are more likely to be observed conditional on the optimal choice not being observed.

An alternative modification of the pure characteristics model is based on adding an idiosyncratic error term to the utility function. This model will have the feature that, conditional on the optimal choice not being observed, a close-to-optimal choice is more likely than a far-from-optimal choice. Suppose the true utility is U_{ij}^* but individuals base their choice on the maximum of mismeasured version of this utility:

$$U_{ij} = U_{ij}^* + \epsilon_{ij},$$

with an extreme value ϵ_{ij}, independent across choices and individuals. The ϵ_{ij} here can be interpreted as an error in the calculation of the utility associated with a particular choice. This model does not directly nest the pure characteristics model, since the idiosyncratic error term has a fixed variance. However, it approximately nests it in the following sense. If the data are generated by the pure characteristics model with the utility function $g(x, \nu, z, \zeta)$, then the model with the utility function $\lambda \cdot g(x, \nu, z, \zeta) + \epsilon_{ij}$ leads, for sufficiently large λ, to choice probabilities that are arbitrarily close to the true choice probabilities (e.g., Berry and Pakes, 2007).

Hence, even if the data were generated by a pure characteristics model, one does not lose much by using a model with an additive idiosyncratic error term, and one gains a substantial amount of robustness to measurement or optimization error.

References

Ackerberg, D., L. Benkard, S. Berry, and A. Pakes, (2005), "Econometric Tools for Analyzing Market Outcomes," forthcoming, *Handbook of Econometrics*, Vol 5, Heckman and Leamer (eds.)

Amemiya, T., and F. Nold, (19750, "A Modified Logit Model," *Review of Economics and Statistics*, Vol 57(2), 255-257.

Athey, S., and G. Imbens, (2008), "Discrete Choice Models with Multiple Unobserved Product Characteristics," *International Economic Review*.

Bajari, P., and L. Benkard, (2004), "Demand Estimation with Heterogenous Consumers and Unobserved Product Characteristics: A Hedonic Approach," Stanford Business School.

Berry, S., (1994), "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, Vol. 25, 242-262.

Berry, S., J. Levinsohn, and A. Pakes, (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, Vol. 63, 841-890.

Berry, S., J. Levinsohn, and A. Pakes (2004), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, Vol 112(1), 68-105.

Berry, S., O. Linton, and A. Pakes, (2004), "Limit Theorems for Estimating the Parameters of Differentiated Product Demand Systems ", *Review of Economic Studies*, Vol. 71, 613-654.

Berry, S., and A. Pakes, (2007), "The Pure Characteristics Discrete Choice Model of Differentiated Products Demand," *International Economic Review*, forthcoming.

Dempster, A., N. Laird, and D. Rubin, (1974), "Maximum Likelihood from Incomplete Data via the EM Algorithm", (with discussion), *Journal of the Royal Statistical*

Society, Series B, Vol. 39, 1-38.

Elrod, T., and M. Keane, (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, "*Journal of Marketing Research*, Vol. XXXII, 1-16.

Geweke, J., M. Keane, and D. Runkle, (1994), "Alternative Computational Approaches to Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 76, No 4, 609-632.

Gill, P., W. Murray, and M. Wright, (1981), *Practical Optimization*, Harcourt Brace and Company, London

Goettler, J., and R. Shachar (2001), "Spatial Competition in the Network Television Industry," *RAND Journal of Economics*, Vol. 32(4), 624-656.

Goldberg, P., (1995), "Product Differentiation and Oligopoly in International Markets: The Case of the Automobile Industry," *Econometrica*, 63, 891-951.

Hajivassiliou, V., and P. Ruud, (1994), "Classical Estimation Methods for LDV Models Using Simulation," in Engle and McFadden 9eds.), *Handbook of Econometrics*, Vol 4, Chapter 40, Elseviers.

Hajivassiliou, V., and D. McFadden, (1990, "The method of simulated scores,", with application to models of external debt crises," unpublished manuscript, Department of Economics, Yale University.

Heckman, J., and B. Singer, (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2).

Manski, C., and S. Lerman,, (1981) "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 305-319, MIT Press, Cambridge, MA.

McCulloch, R., and P. Rossi, (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model, "*Journal of Econometrics* 64 207-240.

McCulloch, R., N. Polson, and P. Rossi, (2000) "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters, "*Journal of Econometrics* 99, 173-193.

McFadden, D., (1973), "Conditional Logit Analysis of Qualitative Choice Behavior "in P. Zarembka (ed), *Frontiers in Econometrics* Academic Press, New York 105-142.

McFadden, D., (1981) "Econometrica Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, Manski and McFadden (eds.), 198-272, MIT Press, Cambridge, MA.

McFadden, D., (1982), "Qualitative Response Models," in Hildenbrand (ed.), *Advances in Econometrics*, Econometric Society Monographs, Cambridge University Press.

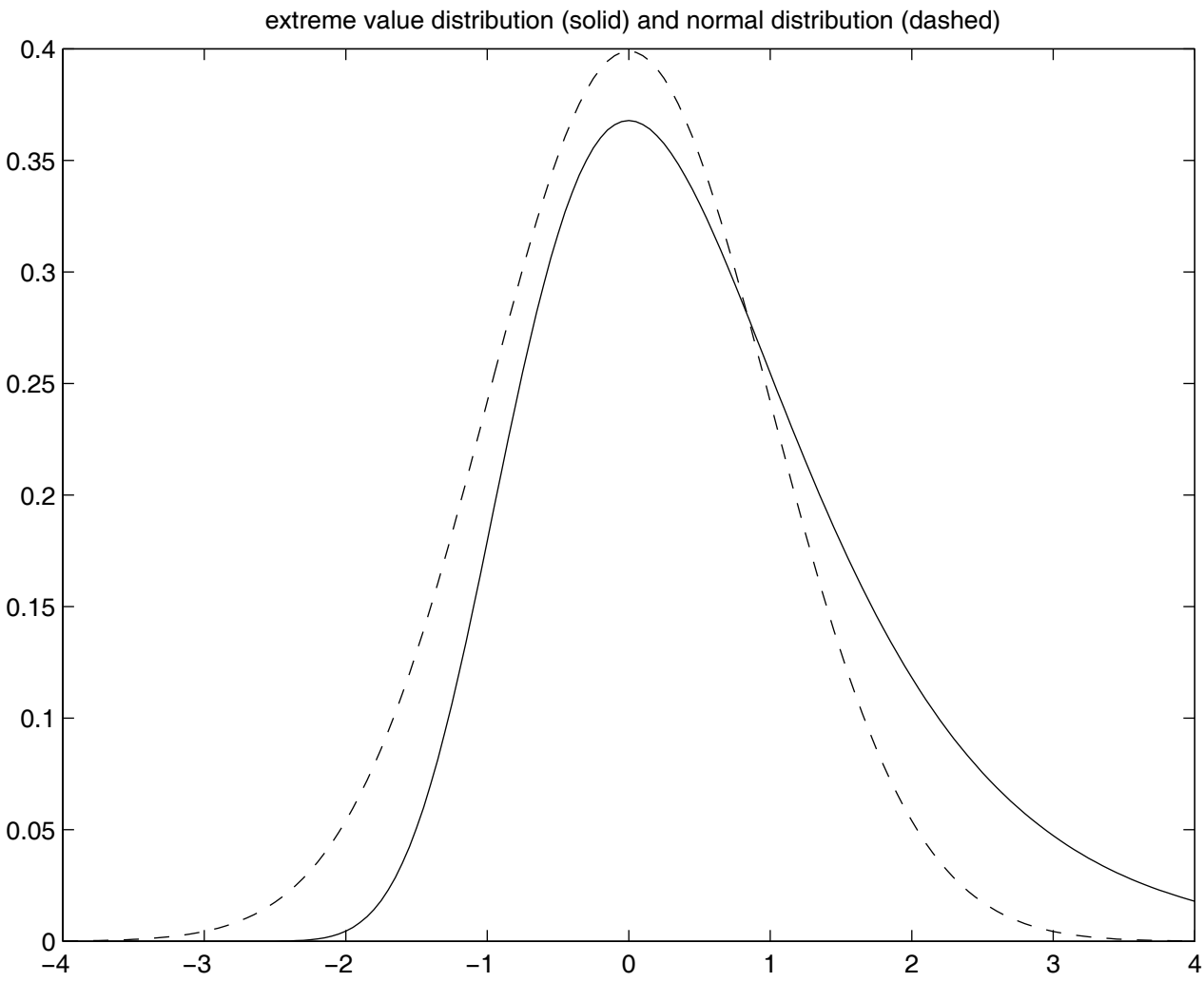
McFadden, D., (1984), "Econometric Analysis of Qualitative Response Models,"in Griliches and Intriligator (eds), *Handbook of Econometrics*, Vol. 2, 1395- 1457, Amsterdam.

McFadden, D., (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 57(5), 995-1026.

McFadden, D., and K. Train, (2000), "Mxed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15(5), 447-470.

Nevo, A. (2000), "A Practioner's Guide to Estimation of Random-Coefficient Logit Models of Demand," *Journal of Economics & Management Science,* Vol. 9, No. 4, 513-548.

Pakes, A., and D. Pollard, (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027-1057.


extreme value distribution (solid) and normal distribution (dashed)
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4
−3
−2
−1
0
1
2
3
4

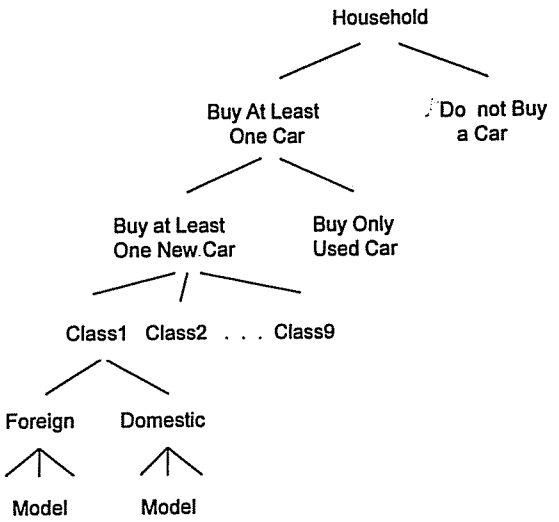


FIGURE 1.—Automobile choice model.

"Cross-Section Econometrics"

Lecture 5

Discrete Choice Models

Guido Imbens

AEA Lectures, Chicago, January 2012

Outline

1. Introduction
2. Multinomial and Conditional Logit Models
3. Independence of Irrelevant Alternatives
4. Models without IIA
5. Berry-Levinsohn-Pakes
6. Models with Multiple Unobserved Choice Characteristics
7. Hedonic Models

1. Introduction

Various versions of multinomial logit models developed by McFadden in 70's.

In IO applications with substantial number of choices IIA property found to be particularly unattractive because of unrealistic implications for substitution patterns.

Random effects approach is more appealing generalization than either nested logit or unrestricted multinomial probit

Generalization by BLP to allow for endogenous choice characteristics, unobserved choice characteristics, using only aggregate choice data.

2. Multinomial and Conditional Logit Models

Models for discrete choice with more than two choices.

The choice Y_i takes on non-negative, unordered integer values between zero and J.

Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan).

We wish to model the distribution of Y in terms of covariates individual-specific, choice-invariant covariates Z_i (e.g., age) choice (and possibly individual) specific covariates X_{ij}.

2.A Multinomial Logit

Individual-specific covariates only.

$$\Pr(Y_i = j|Z_i = z) = \frac{\exp(z'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for choices $j = 1, \ldots, J$ and for the first choice:

$$\Pr(Y_i = 0|Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

The γ_l here are choice-specific parameters. This multinomial logit model leads to a very well-behaved likelihood function, and it is easy to estimate using standard optimization techniques.

2.B Conditional Logit

Suppose all covariates vary by choice (and possibly also by individual). The conditional logit model specifies:

$$\Pr(Y_i = j|X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)},$$

for $j = 0, \ldots, J$. Now the parameter vector β is common to all choices, and the covariates are choice-specific.

Also easy to estimate.

The multinomial logit model can be viewed as a special case of the conditional logit model. Suppose we have a vector of individual characteristics Z_i of dimension K, and J vectors of coefficients γ_j, each of dimension K. Then define

$$X_{i1} = \begin{pmatrix} Z_i \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots\ldots \quad X_{iJ} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ Z_i \end{pmatrix}, \text{ and } X_{i0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and define the parameter vector as $\beta = (\gamma_1', \ldots, \gamma_J')'$. Then

$$\Pr(Y_i = j|Z_i) = \frac{\exp(Z_i'\gamma_j)}{1 + \sum_{k=1}^{J} \exp(Z_i'\gamma_k)}$$

$$= \frac{\exp(X_{ij}'\beta)}{\sum_{k=0}^{J} \exp(X_{ik}'\beta)} = \Pr(Y_i = j|X_{i0}, \ldots, X_{iJ})$$

2.D Link with Utility Maximization

Utility, for individual i, associated with choice j, is

$$U_{ij} = X_{ij}'\beta + \varepsilon_{ij}. \tag{1}$$

i choose option j if choice j provides the highest level of utility

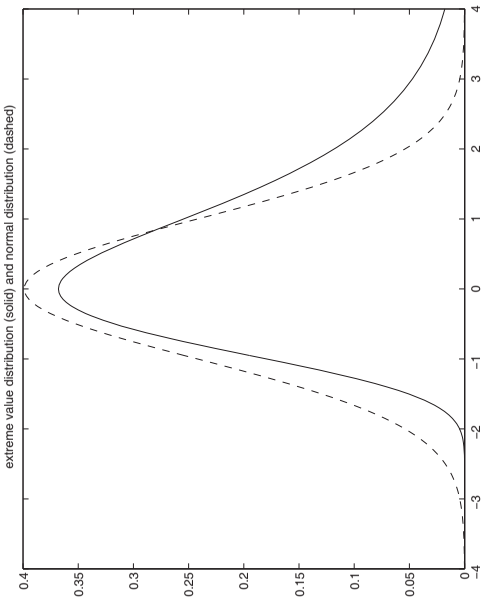
$$Y_i = j \quad \text{if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions.

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

(This distribution has a unique mode at zero, a mean equal to 0.58, and a a second moment of 1.99 and a variance of 1.65.)

Then the choice Y_i follows the conditional logit model.



3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA).

The conditional probability of choosing j given either j or l:

$$\Pr(Y_i = j | Y_i \in \{j, l\}) = \frac{\Pr(Y_i = j)}{\Pr(Y_i = j) + \Pr(Y_i = l)}$$

$$= \frac{\exp(X_{ij}'\beta)}{\exp(X_{ij}'\beta) + \exp(X_{il}'\beta)}.$$

This probability does not depend on the characteristics X_{im} of alternatives m.

Also unattractive implications for marginal probabilities for new choices.

Although multinomial and conditional logit models may fit well, they are not necessarily attractive as behavior/structural models. because they generates unrealistic substitution patterns.

Suppose that individuals have the choice out of three restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose we have two characteristics, price and quality

price	$P_C = 95$, $P_L = 80$, $P_B = 5$,
quality	$Q_C = 10$, $Q_L = 9$, $Q_B = 2$
market share	$S_C = 0.10$, $S_L = 0.25$, $S_B = 0.65$.

These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

$$U_{ij} = -0.2 \cdot P_j + 2 \cdot Q_j + \epsilon_{ij},$$

Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business).

The conditional logit model predicts that the market share for Lalime's gets divided by Chez Panisse and the Bongo Burger, proportional to their original market share, and thus $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$: most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

That seems implausible. The people who were planning to go to Lalime's would appear to be more likely to go to Chez Panisse if Lalime's is closed than to go to the Bongo Burger, implying $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$.

Recall the latent utility set up with the utility

$$U_{ij} = X'_{ij}\beta + \epsilon_{ij}. \tag{2}$$

In the conditional logit model we assume independent extreme value ϵ_{ij}. The independence is essentially what creates the IIA property. (This is not completely correct, because other distributions for the unobserved, say with normal errors, we would not get IIA exactly, but something pretty close to it.)

The solution is to allow in some fashion for correlation between the unobserved components in the latent utility representation. In particular, with a choice set that contains multiple versions of similar choices (like Chez Panisse and LaLime's), we should allow the latent utilities for these choices to be similar.

4. Models without IIA

Here we discuss 3 ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the ϵ_{ij}.

The first is the nested logit model where the researcher groups together sets of choices. This allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation across nests.

Second, the unrestricted multinomial probit model with no restrictions on the covariance between unobserved components, beyond normalizations.

Third, the mixed or random coefficients logit where the marginal utilities associated with choice characteristics vary between individuals, generating positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

Nested Logit Models

Partition the set of choices $\{0, 1, \ldots, J\}$ into S sets $B_1, \ldots, B_S$

Now let the conditional probability of choice j given that your choice is in the set B_s, be equal to

$$\Pr(Y_i = j|X_i, Y_i \in B_s) = \frac{\exp(\rho_s^{-1} X'_{ij}\beta)}{\sum_{l \in B_s} \exp(\rho_s^{-1} X'_{il}\beta)},$$

for $j \in B_s$, and zero otherwise. In addition suppose the marginal probability of a choice in the set B_s is

$$\Pr(Y_i \in B_s|X_i) = \frac{\left(\sum_{l \in B_s} \exp(\rho_s^{-1} X'_{il}\beta)\right)^{\rho_s}}{\sum_{t=1}^{S} \left(\sum_{l \in B_t} \exp(\rho_t^{-1} X'_{il}\beta)\right)^{\rho_s}}.$$

If we fix $\rho_s = 1$ for all s, then

$$\Pr(Y_i = j|X_i) = \frac{\exp(X'_{ij}\beta)}{\sum_{t=1}^{S} \sum_{l \in B_t} \exp(X'_{il}\beta)},$$

and we are back in the conditional logit model.

The implied joint distribution function of the ϵ_{ij} is

$$F(\epsilon_{i0}, \ldots, \epsilon_{iJ}) = \exp\left(-\sum_{s=1}^{S} \left(\sum_{j \in B_s} \exp\left(-\rho_s^{-1}\epsilon_{ij}\right)\right)^{\rho_s}\right).$$

Within the sets the correlation coefficient for the ϵ_{ij} is approximately equal to $1-\rho$. Between the sets the ϵ_{ij} are independent.

The nested logit model could capture the restaurant example by having two nests, the first $B_1 = \{\text{Chez Panisse}, \text{LaLime's}\}$, and the second one $B_2 = \{\text{Bongoburger}\}$.

Estimation of Nested Logit Models

Maximization of the likelihood function is difficult.

An easier alternative is to use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Estimates these as $\widehat{\beta/\rho_s}$.

Then the probability of a particular set B_s can be used to estimate ρ_s through

$$\Pr(Y_i \in B_s|X_i) = \frac{\left(\sum_{l\in B_s} \exp(X'_{il}\widehat{\beta/\rho_s})\right)^{\rho_s}}{\sum_{t=1}^{S}\left(\sum_{l\in B_t} \exp(X'_{il}\widehat{\beta/\rho_t})\right)^{\rho_s}} = \frac{\exp(\rho_s \widehat{W}_s)}{\sum_{t=1}^{S}\exp(\rho_t \widehat{W}_t)},$$

where the "inclusive values" are

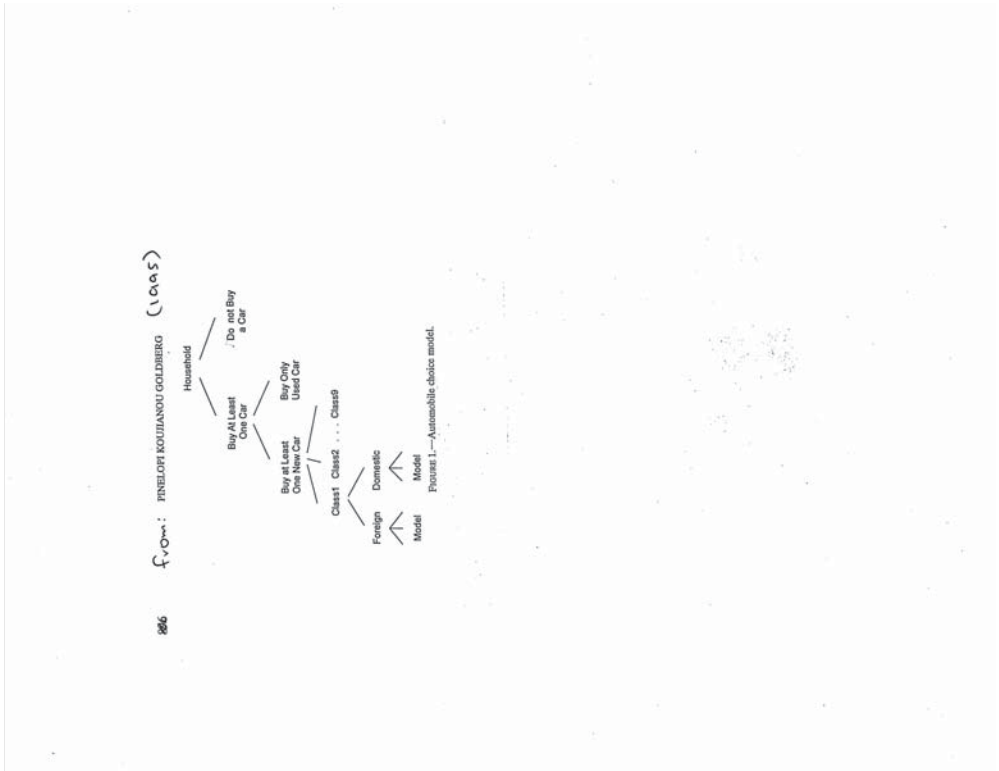
$$\widehat{W}_s = \ln\left(\sum_{l\in B_s} \exp(X'_{il}\widehat{\beta/\rho_s})\right).$$

These models can be extended to many layers of nests. See for an impressive example of a complex model with four layers of multiple nests Goldberg (1995). Figure 2 shows the nests in the Goldberg application.

The key concern with the nested logit models is that results may be sensitive to the specification of the nest structure.

The researcher **chooses** which choices are potentially close substitutes, with the data being used to estimate the amount of correlation.

Researcher would have to choose nest for new good to estimate market share.



Multinomial Probit with Unrestricted Covariance Matrix

A second possibility is to directly free up the covariance matrix of the error terms. This is more natural to do in the multinomial probit case.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X'_{i0}\beta + \epsilon_{i0} \\ X'_{i1}\beta + \epsilon_{i1} \\ \vdots \\ X'_{iJ}\beta + \epsilon_{iJ} \end{pmatrix} \qquad \epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} |X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1)\times(J+1)$ covariance matrix Ω (beyond normalizations).

Direct maximization of the log likelihood function is infeasible for more than 3-4 choices.

Geweke, Keane, and Runkle (1994) and Hajivasilliou and McFadden (1990) proposed a way of calculating the probabilities in the multinomial probit models that allowed researchers to deal with substantially larger choice sets.

A simple attempt to estimate the probabilities would be to draw the ϵ_i from a multivariate normal distribution and calculate the probability of choice j as the number of times choice j corresponded to the highest utility.

The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws $\epsilon_{i1}, \ldots, \epsilon_{iJ}$ sequentially and combines the draws with the calculation of univariate normal integrals.

From a Bayesian perspective drawing from the posterior distribution of β and Ω is straightforward. The key is setting up the vector of unobserved random variables as

$$\theta = (\beta, \Omega, U_{i0}, \ldots, U_{iJ}),$$

and defining the most convenient partition of this vector.

Suppose we know the latent utilities U_i for all individuals. Then the normality makes this a standard linear model problem.

Given the parameters drawing from the unobserved utilities can be done sequentially: for each unobserved utility given the others we would have to draw from a truncated normal distribution, which is straightforward. See McCulloch, Polson, and Rossi (2000) for details.

Merits of Unrestriced Multinomial Probit

The attraction of this approach is that there are no restrictions on which choices are close substitutes.

The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of choices there are a large number of parameters: all elements in the $(J+1) \times (J+1)$ dimensional Ω minus some normalizations and symmetry restrictions.

Estimating all these covariance parameters precisely, based on only first choice data (as opposed to data where we know for each individual additional orderings, e.g., first and second choices), is difficult.

Prediction for new good would require specifying correlations with all other goods.

Random Effects Models

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients.

Why do we fundamentally think that if Lalime's price goes up, the individuals who were planning to go Lalime's go to Chez Panisse instead, rather than to the Bongo Burger? One argument is that we think individuals who have a taste for Lalime's are likely to have a taste for close substitute in terms of observable characteristics, Chez Panisse as well, rather than for the Bongo Burger.

We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X'_{ij}\beta_i + \epsilon_{ij},$$

We can also write this as

$$U_{ij} = X'_{ij}\overline{\beta} + \nu_{ij},$$

where

$$\nu_{ij} = \epsilon_{ij} + X_{ij} \cdot (\beta_i - \overline{\beta}),$$

which is no longer independent across choices.

One possibility to implement this is to assume the existence of a finite number of types of individuals, similar to the finite mixture models used by Heckman and Singer (1984) in duration settings:

$$\beta_i \in \{b_0, b_1, \ldots, b_K\},$$

with

$$\Pr(\beta_i = b_k|Z_i) = p_k, \quad \text{or} \quad \Pr(\beta_i = b_k|Z_i) = \frac{\exp(Z'_i\gamma_k)}{1 + \sum_{l=1}^{K} \exp(Z'_i\gamma_l)}.$$

Here the taste parameters take on a finite number of values, and we have a finite mixture.

Alternatively we could specify

$$\beta_i|Z_i \sim \mathcal{N}(\beta + Z'_i\Gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters.

Using simulation methods or Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

The models with random coefficients can generate more realistic predictions for new choices (predictions will be dependent on presence of similar choices)

5. Berry-Levinsohn-Pakes

BLP extended the random effects logit models to allow for

1. unobserved product characteristics,
2. endogeneity of choice characteristics,
3. estimation with only aggregate choice data
4. with large numbers of choices.

Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products.

The utility is indexed by individual, product and market:

$$U_{ijt} = \beta_i' X_{jt} + \zeta_{jt} + \epsilon_{ijt}.$$

The ζ_{jt} is a unobserved product characteristic. This component is allowed to vary by market and product.

The ϵ_{ijt} unobserved components have extreme value distributions, independent across all individuals i, products j, and markets t.

The random coefficients β_i are related to individual observable characteristics:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i, \quad \text{with} \quad \eta_i | Z_i \sim \mathcal{N}(0, \Sigma).$$

The data consist of

- estimated shares $\widehat{s}_{tj}$ for each choice j in each market t,
- observations from the marginal distribution of individual characteristics (the Z_i's) for each market, often from representative data sets such as the CPS.

First write the latent utilities as

$$U_{ijt} = \delta_{jt} + \nu_{ijt} + \epsilon_{ijt},$$

where

$$\delta_{jt} = \beta' X_{jt} + \zeta_{jt}, \quad \text{and} \quad \nu_{ijt} = (Z_i'\Gamma + \eta_i)' X_{jt}.$$

Now consider for fixed Γ, Σ and δ_{jt} the implied market share for product j in market t, s_{jt}.

This can be calculated analytically in simple cases. For example with $\Gamma_{jt} = 0$ and $\Sigma = 0$, the market share is a very simple function of the δ_{jt}:

$$s_{jt}(\delta_{jt}, \Gamma = 0, \Sigma = 0) = \frac{\exp(\delta_{jt})}{\sum_{l=0}^{J} \exp(\delta_{lt})}.$$

More generally, this is a more complex relationship which we may need to calculate by simulation of choices.

Call the vector function obtained by stacking these functions for all products and markets $s(\delta, \Gamma, \Sigma)$.

Next, fix only Γ and Σ. For each value of δ_{jt} we can find the implied market share. Now find the vector of δ_{jt} such that all implied market shares are equal to the observed market shares $\widehat{s}_{jt}$.

BLP suggest using the following algorithm. Given a starting value for δ_{jt}^0, use the updating formula:

$$\delta_{jt}^{k+1} = \delta_{jt}^k + \ln s_{jt} - \ln s_{jt}(\delta^k, \Gamma, \Sigma).$$

BLP show this is a contraction mapping, and so it defines a function $\delta(s, \Gamma, \Sigma)$ expressing the δ as a function of observed market shares s, and parameters Γ and Σ.

Given this function $\delta(s, \Gamma, \Sigma)$ define the residuals

$$\omega_{jt} = \delta_{jt}(s, \Gamma, \Sigma) - \beta' X_{jt}.$$

At the true values of the parameters and the true market shares these residuals are equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM given instruments that are orthogonal to these residuals, typically things like characteristics of other products by the same firm, or average characteristics by competing products.

This step is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

Let us see what this does if we have, and know we have, a conditional logit model with fixed coefficients. In that case $\Gamma = 0$, and $\Sigma = 0$. Then we can invert the market share equation to get the market specific unobserved choice-characteristics

$$\delta_{jt} = \ln s_{jt} - \ln s_{0t},$$

where we set $\delta_{0t} = 0$. (this is typically the outside good, whose average utility is normalized to zero). The residual is

$$\zeta_{jt} = \delta_{jt} - \beta' X_{jt} = \ln s_{jt} - \ln s_{0t} - \beta' X_{jt}.$$

With a set of instruments W_{jt}, we run the regression

$$\ln s_{jt} - \ln s_{0t} = \beta' X_{jt} + \epsilon_{jt},$$

using W_{jt} as instrument for X_{jt}, using as the observational unit the market share for product j in market t.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach can allow only for a single unobserved choice characteristic. This is essential for their estimation strategy with aggregate data.

With individual level data one may be able to establish the presence of two unobserved product characteristics (invariant across markets). Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007) study such models.

These models can be viewed as freeing up the covariance matrix of unobserved components relative to the random coefficients model, but using a factor structure instead of a fully unrestricted covariance matrix as in the multinomial probit.

Athey and Imbens model the latent utility for individual i in market t for choice j as

$$U_{ijt} = X'_{it}\beta_i + \zeta'_j\gamma_i + \epsilon_{ijt},$$

with the individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Gamma Z_i, \Sigma).$$

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult (multiple modes). Bayesian methods, and in particular markov-chain-monte-carlo methods are more effective tools for conducting inference in these settings.

7. Hedonic Models

Recently researchers have reconsidered using pure characteristics models for discrete choices, that is models with no idiosyncratic error ϵ_{ij}, instead relying solely on the presence of a small number of unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices.

Why can it still be useful to include such an ϵ_{ij}?

First, the pure characteristics model can be extremely sensitive to measurement error, because it can predict zero market shares for some products.

Consider a case where choices are generated by a pure characteristics model that implies that a particular choice j has zero market share. Now suppose that there is a single unit i for whom we observe, due to measurement error, the choice $Y_i = j$.

Irrespective of the number of correctly measured observations available that were generated by the pure characteristics model, the estimates of the latent utility function will not be close to the true values due to a **single** mismeasured observation.

Thus, one might wish to generalize the model to be more robust. One possibility is to related the observed choice Y_i to the optimal choice Y_i^*:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J)$$

$$= \begin{cases} 1-\delta & \text{if } Y_i^* = y, \\ \delta/(J-1) & \text{if } Y_i^* \neq y. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without the extreme sensitivity.

However, if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely.

An alternative modification of the pure characteristics model is based on adding an idiosyncratic error term to the utility function. This model will have the feature that, conditional on the optimal choice not being observed, a close-to-optimal choice is more likely than a far-from-optimal choice.

Suppose the true utility is U_{ij}^* but individuals base their choice on the maximum of mismeasured version of this utility:

$$U_{ij} = U_{ij}^* + \epsilon_{ij},$$

with an extreme value ϵ_{ij}, independent across choices and individuals. The ϵ_{ij} here can be interpreted as an error in the calculation of the utility associated with a particular choice.

Second, this model approximately nests the pure characteristics model in the following sense. If the data are generated by the pure characteristics model with the utility function $g(x, \nu, z, \zeta)$, then the model with the utility function $\lambda \cdot g(x, \nu, z, \zeta) + \epsilon_{ij}$ leads, for sufficiently large λ, to choice probabilities that are arbitrarily close to the true choice probabilities (e.g., Berry and Pakes, 2007).

Hence, even if the data were generated by a pure characteristics model, one does not lose much by using a model with an additive idiosyncratic error term, and one gains a substantial amount of robustness to measurement or optimization error.

AEA Lectures **Chicago, IL, January 2012**

Lecture 7a, Wednesday, Jan 7th, 8.00am-9.45am

Weak Instruments and Many Instruments

1. Introduction

In recent years a literature has emerged that has raised concerns with the quality of inferences based on conventional methods such as Two Stage Least Squares (TSLS) and Limited Information Maximum Likelihood (LIML) in instrumental variables settings when the instrument(s) is/are only weakly correlated with the endogenous regressor(s). Although earlier work had already established the poor quality of conventional normal approximations with weak or irrelevant instruments, the recent literature has been motivated by empirical work where *ex post* conventional large sample approximations were found to be misleading. The recent literature has aimed at developing better estimators and more reliable methods for inference.

There are two aspects of the problem. In the just-identified case (with the number of instruments equal to the number of endogenous regressors), or with low degrees of over-identification, the focus has largely been on the construction of confidence intervals that have good coverage properties even if the instruments are weak. Even with very weak, or completely irrelevant, instruments, conventional methods are rarely substantively misleading, unless the degree of endogeneity is higher than one typically encounters in studies using cross-section data. Conventional TSLS or LIML confidence intervals tend to be wide when the instrument is very weak, even if those intervals do not have the correct nominal coverage for all parts of the parameter space. In this case better estimators are generally not available. Improved methods for confidence intervals based on inverting test statistics have been developed although these do not have the simple form of an estimate plus or minus a constant times a standard error.

The second case of interest is that with a high degree of over-identification. These settings often arise by interacting a set of basic instruments with exogenous covariates in order to

improve precision. If there are many (weak) instruments, standard estimators can be severely biased, and conventional methods for inference can be misleading. In particular TSLS has been found to have very poor properties in these settings. Bootstrapping does not solve these problems. LIML is generally much better, although conventional LIML standard errors are too small. A simple to implement proportional adjustment to the LIML standard errors based on the Bekker many-instrument asymptotics or the Chamberlain-Imbens random coefficients argument appears to lead to substantial improvements in coverage rates.

2. Motivation

Much of the recent literature is motivated by a study by Angrist and Krueger (1991, AK). Subsequently Bound, Jaeger and Baker (1996, BJB) showed that for some specifications AK employed normal approximations were not appropriate despite very large sample sizes (over 300,000 observations).

2.1 The Angrist-Krueger Study

AK were interested in estimating the returns to years of education. Their basic specification is:

$$Y_i = \alpha + \beta \cdot E_i + \varepsilon_i,$$

where Y_i is log (yearly) earnings and E_i is years of education. Their concern, following a long literature in economics, e.g., Griliches, (1977), Card (2001), is that years of schooling may be endogenous, with pre-schooling levels of ability affecting both schooling choices and earnings given education levels. In an ingenuous attemp to address the endogeneity problem AK exploit variation in schooling levels that arise from differential impacts of compulsory schooling laws. School districts typically require a student to have turned six by January 1st of the year the student enters school. Since individuals are required to stay in school till they turn sixteen, individual born in the first quarter have lower required minimum schooling levels than individuals born in the last quarter. The cutoff dates and minimum school dropout age differ a little bit by state and over time, so the full picture is more complicated but the basic point is that the compulsory schooling laws generate variation in

schooling levels by quarter of birth that AK exploit. Let Q_i be the indicator for being born in the fourth quarter.

One can argue that a more natural analysis of such data would be as a Regression Discontinuity (RD) design, where we focus on comparisons of individuals born close to the cutoff date. We will discuss such designs in a later lecture. However, in the census only quarter of birth is observed, not the actual date, so there is in fact little that can be done with the RD approach beyond what AK do. In addition, there are substantive arguments why quarter of birth need not be a valid instrument (e.g., seasonal patterns in births, or differential impacts of education by age at entering school). AK discuss many of the potential concerns. See also Bound, Jaeger and Baker (1996). We do not discuss these concerns here further.

Table 1 shows average years of education and average log earnings for individual born in the first and fourth quarter, using the 1990 census. This is a subset of the AK data.

Table 1: Summary Statistics Subset of AK Data

Variable	1st Quarter	4th Quarter	difference
Year of Education	12.688	12.840	0.151
Log Earnings	5.892	5.905	0.014
ratio			0.089

The sample size is 162,487. The last column gives the difference between the averages by quarter, and the last row the ratio of the difference in averages. The last number is the Wald estimate of the returns to education based on these data:

$$\hat{\beta} = \frac{\overline{Y}_4 - \overline{Y}_1}{\overline{E}_4 - \overline{E}_1} = 0.0893 \quad (0.0105),$$

where $\overline{Y}_t$ and $\overline{E}_t$ are the average level of log earnings and years of education for individuals born in the t-th quarter. This is also equal to the Two-Stage-Least-Squares (TSLS) and

Limited-Information-Maximum-Likelihood (LIML) estimates because there is only a single instrument and a single endogenous regressor. The standard error here is based on the delta method and asymptotic joint normality of the numerator and denominator.

AK also present estimates based on additional instruments. They take the basic instrument and interact it with 50 state and 9 year of birth dummies. Here we take this a bit further, and following Chamberlain and Imbens (2004) we interact the single binary instrument with state times year of birth dummies to get 500 instruments. Denote the 500 dimensional vector of interactions of year and state of birth by W_i, and let $X_i = (W_i' E_i)'$ be 501 dimensional the vector of included covariates (both endogenous and exogenous) and $Z_i = (W_i', Q_i \cdot W_i')'$ be the 1000 dimensional vector of exogenous variables (including both the excluded instruments $Q_i \cdot W_i$ and the included exogenosu regressors W_i). This leads to the following model:

$$Y_i = X_i'\beta + \varepsilon_i = W_i'\beta_0 + E_i \cdot \beta_1 + \varepsilon_i, \qquad \mathbb{E}[Z_i \cdot \varepsilon_i] = 0.$$

Let $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ be the $N \times 1$ vector of log earnings, the $N \times 501$ matrix with regressors, and the $N \times 1000$ matrix of instruments. The TSLS estimator for β is then

$$\hat{\beta}_{\text{TSLS}} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\right).$$

For these data this leads to

$$\hat{\beta}_{\text{TSLS}} = 0.073 \quad (0.008).$$

The LIML estimator adds a normal model for the relation between the L vector X_i and the K-vector Z_i, of the form

$$E_i = \pi' Z_i + \nu_i,$$

and is based on maximization of the log likelihood function

$$L(\beta, \pi, \Omega) = \sum_{i=1}^{N}\left(-\frac{1}{2}\ln|\Omega| - \frac{1}{2}\begin{pmatrix} Y_i - \beta_0' W_i - \beta_1 \cdot \pi' Z_i \\ E_i - \pi' Z_i \end{pmatrix}' \Omega^{-1} \begin{pmatrix} Y_i - \beta_0' W_i - \beta_1 \cdot \pi' Z_i \\ E_i - \pi' Z_i \end{pmatrix}\right),$$

where Ω is the reduced form covariance matrix (the covariance matrix of $(\varepsilon, \nu_i)'$.

For this subset of the AK data we find, for the coefficient on years of education,

$$\hat{\beta}_{\text{LIML}} = 0.095 \quad (0.017).$$

In large samples the LIML and TSLS are equivalent under homoskedasticity.

2.2 The Bound-Jaeger-Baker critique

BJB found that are potential problems with the AK results. They suggested that despite the large samples used by AK large sample normal approximations may be very poor. The reason is that the instruments are only very weakly correlated with the endogenous regressor. The most striking evidence for this is based on the following calculations, that are based on a suggestion by Alan Krueger. Take the AK data and re-calculate their estimates after replacing the actual quarter of birth dummies by random indicators with the same marginal distribution. In principle this means that the standard (gaussian) large sample approximations for TSLS and LIML are invalid since they rely on non-zero correlations between the instruments and the endogenous regressor. Doing these calculations once for the single and 500 instrument case, for both TSLS and LIML, leads to the results in Table 2

Table 2: Real and Random QOB Estimates

	Single Instrument		500 Instruments			
			TSLS		LIML	
Real QOB	0.089	(0.011)	0.073	(0.008)	0.095	(0.017) [0.037]
Random QOB	0.181	(0.193)	0.059	(0.009)	-0.134	(0.065) [0.251]

With the single instrument the results are not so disconcertening. Although the confidence interval is obviously not valid, it is wide, and few researchers would be misled by the results. With many instruments the results are much more troubling. Although the instrument con-

tains no information, the results suggest that the instruments can be used to infer precisely what the returns to education are. These results have provided the motivation for the recent weak instrument literature. Note that there is an earlier literature, e.g., Phillips (1984) Rothenberg (1984), but it is the BJB findings that got the attention of researchers doing empirical work.

2.3 Simulations with Weak Instruments and Varying Degrees of Endogeneity

Here we provide slightly more systematic simulation evidence of the weak instrument problems in the AK setting. We create 10,000 artificial data sets, all of size 160,000, designed to mimic the key features of the AK data. In each of these data sets half the units have quarter of birth (denoted by Q_i) equal to 0 and 1 respectively. Then we draw the two reduced form residuals ν_i and η_i from a joint normal distribution

$$\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.446 & \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} \\ \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} & 10.071 \end{pmatrix} \right).$$

The variances of the reduced form errors mimic those in the AK data. The correlation between the reduced form residuals in the AK data is 0.318. The implied OLS coefficient is $\rho \cdot \sqrt{0.446}/\sqrt{10.071}$. Then years of education is equal to

$$E_i = 12.688 + 0.151 \cdot Q_i + \eta_i,$$

and log earnings is equal to

$$Y_i = 5.892 + 0.014 \cdot Q_i + \nu_i.$$

Now we calculate the IV estimator and its standard error, using either the actual qob variable or a random qob variable as the instrument. We are interested in the size of tests of the null that coefficient on years of education is equal to $0.089 = 0.014/0.151$. We base the test on the t-statistic. Thus we reject the null if the ratio of the point estimate minus 0.089 and the standard error is greater than 1.96 in absolute value. We repeat this for 12 different values of the reduced form error correlation. In Table 3 we report the proportion of rejections and the median and 0.10 quantile of the width of the estimated 95% confidence intervals.

Table 3: Coverage Rates of Conv. TSLS CI by Degree of Endogeneity

ρ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
implied OLS	0.00	0.02	0.04	0.06	0.08	0.11	0.13	0.15	0.17	0.19	0.20	0.21
Real QOB	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.95	0.95	0.95	0.95
Med Width 95% CI	0.09	0.09	0.09	0.08	0.08	0.08	0.07	0.07	0.06	0.05	0.05	0.05
0.10 quant Width	0.08	0.08	0.08	0.07	0.07	0.07	0.06	0.06	0.05	0.04	0.04	0.04
Random QOB	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.92	0.82	0.53
Med Width 95% CI	1.82	1.81	1.78	1.73	1.66	1.57	1.45	1.30	1.09	0.79	0.57	0.26
0.10 quant Width	0.55	0.55	0.5403	0.53	0.51	0.48	0.42	0.40	0.33	0.24	0.17	0.08

In this example, unless the reduced form correlations are very high, e.g., at least 0.95, with irrelevant the conventional confidence intervals are wide and have good coverage. The amount of endogeneity that would be required for the conventional confidence intervals to be misleading is higher than one typically encounters in cross-section settings. It is likely that these results extend to cases with a low degree of over-identification, using either TSLS, or preferably LIML. Put differently, although formally conventional confidence intervals are not valid uniformly over the parameter space (e.g., Dufour, 1997), there are no examples we are aware of where they have substantively misleading in just-identified examples. This in contrast to the case with many weak instruments where especially TSLS can be misleading in empirically relevant settings.

3. Weak Instruments

Here we discuss the weak instrument problem in the case of a single instrument, a single endogenous regressor, and no additional exogenous regressors beyond the intercept. More generally the qualitative features of these results by and large apply to the case with a few

weak instruments. We consider the model

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i,$$

$$X_i = \pi_0 + \pi_1 \cdot Z_i + \eta_i,$$

with $(\varepsilon_i, \eta_i) \perp\!\!\!\perp Z_i$, and jointly normal with covariance matrix Σ. (The normality is mainly for some of the exact results, and it does not play an important role.) The reduced form for the first equation is

$$Y_i = \alpha_0 + \alpha_1 \cdot Z_i + \nu_i,$$

where the parameter of interest is $\beta_1 = \alpha_1/\pi_1$. Let

$$\Omega = \mathbb{E}\left[\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix}' \right], \qquad \text{and} \quad \Sigma = \mathbb{E}\left[\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}' \right],$$

be the covariance matrix of the reduced form and stuctural disturbances respectively. Many of the formal results in the literature are for the case of known Ω, and normal disturbances. This is largely innocuous, as Ω can be precisely estimated in typical data sets. Note that this it not the same as assuming that Σ is known, which is not innocuous since it depends on Ω and β, and cannot be precisely estimated in settings with weak instruments

$$\Sigma = \begin{pmatrix} \Omega_{11} - 2\beta\Omega_{12} + \beta^2\Omega_{22} & \Omega_{12} - \beta\Omega_{22} \\ \Omega_{12} - \beta\Omega_{22} & \Omega_{22} \end{pmatrix}.$$

The standard estimator for β_1 is

$$\hat{\beta}_1^{\text{IV}} = \frac{\frac{1}{N}\sum_{i=1}^N \left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right)}{\frac{1}{N}\sum_{i=1}^N \left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right)},$$

where $\overline{Y} = \sum_i Y_i/N$, and similarly for $\overline{X}$ and $\overline{Z}$.

A simple interpretation of the weak instrument is that with the concentration parameter

$$\lambda = \pi_1^2 \cdot \sum_{i=1}^N \left(Z_i - \overline{Z}\right)^2 \Big/ \sigma_\eta^2.$$

close to zero, both the covariance in the numerator and the covariance in the denominator are close to zero. In reasonably large samples both are well approximated by normal

distributions:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)\left(Z_i-\overline{Z}\right)-\mathrm{Cov}(Y_i,Z_i)\right)\approx\mathcal{N}\left(0,V(Y_i\cdot Z_i)\right),$$

and

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Z_i-\overline{Z}\right)-\mathrm{Cov}(X_i,Z_I)\right)\approx\mathcal{N}\left(0,V(X_i\cdot Z_i)\right).$$

These two normal approximations tend to be accurate in applications with reasonable sample sizes, irrespective of the population values of the covariances. If $\pi_1 \neq 0$, as the sample size gets large, then the ratio will eventually be well approximated by a normal distribution as well. However, if $\mathrm{Cov}(X_i, Z_i) \approx 0$, the ratio may be better approximated by a Cauchy distribution, as the ratio of two normals centered close to zero.

The weak instrument literature is concerned with inference for β_1 when the concentration parameter λ is too close to zero for the normal approximation to the ratio to be accurate.

Staiger and Stock (1997, SS) formalize the problem by investigating the distribution of the standard IV estimator under an alternative asymptotic approximation. The standard asymptotics (strong instrument asymptotics in the SS terminology) is based on fixed parameters and the sample size getting large. In their alternative asymptotic sequence SS model π_1 as a function of the sample size, $\pi_{1N} = c/\sqrt{N}$, so that the concentration parameter converges to a constant:

$$\lambda \longrightarrow c^2 \cdot V(Z_i).$$

SS then compare coverage properties of various confidence intervals under this (weak instrument) asymptotic sequence.

The importance of the SS approach is not in the specific sequence. The concern is more that if a particular confidence interval does not have the appropriate coverage asymptotically under the SS asymptotics, then there are values of the (nuisance) parameters in a potentially important part of the parameter space (namely around $\pi_i = 0$) such that the actual coverage is substantially away from the nominal coverage for any sample size. More recently the issue

has therefore been reformulated as requiring confidence intervals to have asymptotically the correct coverage probabilities uniformly in the parameter space. See for a discussion from this perspective Mikusheva (2007). For estimation this perspective is not helpful: there cannot be estimators that are consistent for β^* uniformly in the parameter space since if $\pi_1 = 0$, there are no consistent estimators for β_1. However, for testing there are generally confidence intervals that are uniformly valid, but they are not of the conventional form, that is, a point estimate plus or minus a constant times a standard error.

3.1 Tests and Confidence Intervals in the Just-identified Case

Let the instrument $\tilde{Z}_i = Z_i - \overline{Z}$ be measured in deviations from its mean. Then define the statistic

$$S(\beta_1) = \frac{1}{N}\sum_{i=1}^{N} \tilde{Z}_i \cdot (Y_i - \beta_1 \cdot X_i)\,.$$

Then, under the null hypothesis that $\beta_1 = \beta_1^*$, and conditional on the instruments, the statistic $\sqrt{N} \cdot S(\beta_1^*)$ has an exact normal distribution

$$\sqrt{N} \cdot S(\beta_1^*) \sim \mathcal{N}\left(0, \sum_{i=1}^{N} \tilde{Z}_i^2 \cdot \sigma_\varepsilon^2\right).$$

Importantly, this result does not depend on the strength of the instrument. Anderson and Rubin (1949, AR) propose basing tests for the null hypothesis

$$H_0: \quad \beta_1 = \beta_1^0, \qquad \text{against the alternative hypothesis} \quad H_a: \quad \beta_1 \neq \beta_1^0,$$

on this idea, through the statistic

$$\text{AR}\left(\beta_1^0\right) = \frac{N \cdot S(\beta_1^0)^2}{\sum_{i=1}^{N} \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix} \right)^{-1}.$$

This statistic has an exact chi-squared distribution with degrees of freedom equal to one. In practice, of course, one does not know the reduced form covariance matrix Ω, but substituting an estimated version of this matrix based on the average of the estimated reduced form residuals does not affect the large sample properties of the test.

A confidence interval can be based on this test statistic by inverting it. For example, for a 95% confidence interval for β_1, we would get

$$\text{CI}^{\beta_1}_{0.95} = \{\beta_1 \,|\text{AR}(\beta_1) \leq 3.84\}\,.$$

Note that this AR confidence interval cannot be empty, because at the standard IV estimator $\hat{\beta}_1^{\text{IV}}$ we have $\text{AR}(\hat{\beta}_1^{\text{IV}}) = 0$, and thus $\hat{\beta}_1^{\text{IV}}$ is always in the confidence interval. The confidence interval can be equal to the entire real line, if the correlation between the endogenous regressor and the instrument is close to zero. This is not surprising: in order to be valid even if $\pi_1 = 0$, the confidence interval must include all real values with probability 0.95.

3.3 Tests and Confidence Intervals in the Over-identified Case

The second case of interest is that with a single endogenous regressor and multiple instruments. We deal separately with the case where there are many (similar) instrument, so this really concerns the case where the instruments are qualitatively different. Let the number of instrumens be equal to K, so that the reduced form is

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i,$$

with Z_i a k-dimensional column vector. There is still only a single endogenous regressor, and no exogenous regressors beyond the intercept. All the results generalize to the case with additional exogenous covariates at the expense of additional notatio. The AR approach can be extended easily to this over-identified case, because the statistic $\sqrt{N}\cdot S(\beta_1^*)$ still has a normal distribution, but now a multivariate normal distribution. Hence one can base tests on the AR statistic

$$\text{AR}\left(\beta_1^0\right) = N\cdot S\left(\beta_1^0\right)'\left(\sum_{i=1}^N \tilde{Z}_i\cdot\tilde{Z}_i'\right)^{-1} S\left(\beta_1^0\right)\cdot\left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix}\Omega\begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

Under the same conditions as before this has an exact chi-squared distribution with degrees of freedom equal to the number of instruments, k. A practical problem arises if we wish to construct confidence intervals based on this statistic. Suppose we construct a confidence interval, analogously to the just-identified case, as

$$\text{CI}^{\beta_1}_{0.95} = \left\{\beta_1 \,\middle|\text{AR}(\beta_1) \leq \mathcal{X}^2_{0.95}(K)\right\},$$

where $\mathcal{X}^2_{0.95}(k)$ is the 0.95 quantile of the chi-squared distribution with degrees of freedom equal to k. The problem is that this confidence interval can be empty. The interpretation is that the test does not only test whether $\beta_1 = \beta_1^0$, but also tests whether the instruments are valid. However, one generally may not want to combine those hypotheses.

Kleibergen (2002) modifies the AR statistic and confidence interval construction. Instead of the statistic $S(\beta_1)$, he considers a statistic that looks at the correlation between a particular linear combination of the instruments (namely the estimated endogenous regressor) and the residual:

$$\tilde{S}\left(\beta_1^0\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Z}_i'\hat{\pi}_1(\beta_1^0)\right)\cdot\left(Y_i - \beta_1^0\cdot X_i\right),$$

where $\hat{\pi}$ is the maximum likelihood estimator for π_1 under the restriction $\beta_1 = \beta_1^0$. The test is then based on the statistic

$$K\left(\beta_1^0\right) = \frac{N\cdot S(\beta_1^0)^2}{\sum_{i=1}^N \tilde{Z}_i^2}\cdot\left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix}\Omega\begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

This statistic has no longer an exact chi-squared distribution, but in large samples it still has an approximate chi-square distribution with degrees of freedom equal to one. Hence the test is straightforward to implement using standard methods.

Moreira (2003) proposes a method for adjusting the critical values that applies to a number of tests, including the Kleibergen test. His idea is to focus on *similar* tests, test that have the same rejection probability for all values of the nuisance parameter. The nuisance parameter is here the vector of reduced form coefficients π, since we assume the residual covariance matrix is known. The way to adjust the critical values is to consider the distribution of a statistic such as the Kleibergen statistic conditional on a complete sufficient statistic for the nuisance parameter. In this setting a complete sufficient statistic is readily available in the form of the maximum likelihood estimator under the null, $\hat{\pi}_1(\beta_1^0)$. Moreira's preferred test is based on the likelihood ratio. Let

$$LR\left(\beta_1^0\right) = 2\cdot\left(L\left(\hat{\beta}_1,\hat{\pi}\right) - L\left(\beta_1^0,\hat{\pi}(\beta_1^0)\right)\right),$$

be the likelihood ratio. Then let $c_{LR}(p, 0.95)$, be the 0.95 quantile of the distribution of $LR(\beta_1^0)$ under the null hypothesis, conditional on $\hat{\pi}(\beta_1^0) = p$. The proposed test is to reject the null hypothesis at the 5% level if

$$LR\left(\beta_1^0\right) > c_{LR}(\hat{\pi}(\beta_1^0), 0.95),$$

where conventional test would use critical values from a chi-squared distribution with a single degree of freedom. This test can then be converted to construct a 95% confidence intervals. Calculation of the (large sample) critical values is simplified by the fact that they only depend on the number of instruments k, and a scaled version of the $\hat{\pi}(\beta_1^0)$. Tabulations of these critical values are in Moreira (2003) and have been programmed in STATA (See Moreira's website).

3.4 Conditioning on the First Stage

The AR, Kleibergen and Moreira proposals for confidence intervals are asymptotically valid irrespective of the strength of the first stage (the value of π_1). However, they are not valid if one first inspects the first stage, and conditional on the strength of that, decides to proceed. Specifically, if in practice one first inspects the first stage, and decide to abandon the project if the first stage F-statistic is less than some fixed value, and otherwise proceed by calculating an AR, Kleibergen or Moreira confidence interval, the large sample coverage probabilities would not necessarily be the nominal ones. In practice researchers do tend to inspect and report the strength of the first stage. This is particularly true in recent instrumental variables literature where researchers argue extensively for the validity of the instrumental variables assumption. This typically involves detailed arguments supporting the alleged mechanism that leads to the correlation between the endogenous regressor and the instruments. For example, Section I in AK (page 981-994) is entirely devoted to discussing the reasons and evidence for the relation between their instruments (quarter of birth) and years of education. In such cases inference conditional on this may be more appropriate.

Chioda and Jansson (2006) propose a clever alternative way to construct a confidence interval that is valid conditional on the strength of the first stage. Their proposed confidence

interval is based on inverting a test statistic similar to the AR statistic. It has a non-standard distribution conditional on the strength of the first stage, and they suggest a procedure that involves numerically approximating the critical values. A caveat is that because the first stage F-statistic, or the first stage estimates are not ancillary, conditioning on them involves loss of information, and as a result the Chioda-Jansson confidence intervals are wider than confidence intervals that are not valid conditional on the first stage.

4. Many Weak Instruments

In this section we discuss the case with many weak instruments. The problem is both the bias in the standard estimators, and the misleadingly small standard errors based on conventional procedures, leading to poor coverage rates for standard confidence intervals in many situations. The earlier simulations showed that especially TSLS, and to a much lesser extent LIML, have poor properties in this case. Note first that resampling methods such as bootstrapping do not solve these problems. In fact, if one uses the standard bootstrap with TSLS in the AK data, one finds that the average of the bootstrap estimates is very close to the TSLS point estimat, and that the bootstrap variance is very close to the TSLS variance.

The literature has taken a number of approaches. Part of the literature has focused on alternative confidence intervals analogues to the single instrument case. In addition a variety of new point estimators have been proposed.

4.1 Bekker Asymptotics

In this setting alternative asymptotic approximations play a bigger role than in the single instrument case. In an important paper Bekker (1995) derives large sample approximations for TSLS and LIML based on sequences where the number of instruments increases proportionally to the sample size. He shows that TSLS is not consistent in that case. LIML is consistent, but the conventional LIML standard errors are not valid. Bekker then provides LIML standard errors that are valid under this asymptotic sequence. Even with relatively small numbers of instruments the differences between the Bekker and conventional asymptotics can be substantial. See also Chao and Swanson (2005), and Hansen, Hausman and

Newey () for extensions.

Here we describe the Bekker correction to the standard errors for the model with a single endogenous regressors, allowing for the presence of exogenous regressors. We write the model as:

$$Y_i = \beta_1' X_{1i} + \beta_2' X_{2i} + \varepsilon_i = \beta' X_i + \varepsilon_i,$$

where the single endogenous variable X_{1i} satisfies:

$$X_{1i} = \pi_1' Z_{1i} + \pi_2' X_{2i} + \eta_i = \pi' Z_i + \eta_i.$$

Define the matrices $\mathbf{P}_Z$ and $\mathbf{M}_Z$ as:

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}', \qquad \mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

Let σ^2 be the variance of ε_i, with consistent estimator $\hat{\sigma}^2$. The standard TSLS variance is

$$V_{tsls} = \hat{\sigma}^2 \cdot \left(\mathbf{X}\mathbf{P}_Z\mathbf{X}\right)^{-1}.$$

Under the standard, fixed number of instrument asymptotics, the asymptotic variance for LIML is identical to that for TSLS, and so in principle we can use the same estimator. In practice researchers typically estimate the variance for LIML as

$$V_{liml} = \hat{\sigma}^2 \cdot \Big(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda} \cdot \mathbf{X}'\mathbf{M}_Z\mathbf{X}\Big)^{-1},$$

To get Bekker's correction, we need a little more notation. Define

$$\Omega = \left(\begin{array}{cc} \mathbf{Y} & \mathbf{X} \end{array} \right) \mathbf{P}_Z \left(\begin{array}{cc} \mathbf{Y} & \mathbf{X} \end{array} \right) / N = \left(\begin{array}{cc} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{array} \right),$$

so that

$$\Omega_{11} = \mathbf{Y}\mathbf{P}_Z\mathbf{Y}/N, \qquad \Omega_{12} = \mathbf{Y}\mathbf{P}_Z\mathbf{X}/N, \qquad \text{and} \quad \Omega_{22} = \mathbf{X}\mathbf{P}_Z\mathbf{X}/N.$$

Now define

$$\mathbf{A} = N \cdot \frac{\Omega_{12}'\Omega_{12} - \Omega_{22}\beta\Omega_{12} - \Omega_{12}'\beta'\Omega_{22} + \Omega_{22}\beta\beta'\Omega_{22}}{\Omega_{11} - 2\Omega_{12}\beta + \beta'\Omega_{22}\beta}.$$

Then:

$$V_{bekker} = \hat{\sigma}^2 \cdot \Big(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda}\cdot\mathbf{X}'\mathbf{M}_Z\mathbf{X}\Big)^{-1}$$
$$\times \left(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \lambda\cdot\mathbf{A}\right)\cdot\Big(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda}\cdot\mathbf{X}'\mathbf{M}_Z\mathbf{X}\Big)^{-1}.$$

4.2 Random Effects Estimators

Chamberlain and Imbens (2004, CI) propose a random effects quasi maximum likelihood estimator. They propose modelling the first stage coefficients π_k, for $k = 1, \ldots, K$, in the regression

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i = \pi_0 + \sum_{k=1}^{K} \pi_k \cdot Z_{ik} + \eta_i,$$

(after normalizing the instruments to have mean zero and unit variance,) as independent draws from a normal $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ distribution. (More generally CI allow for the possibility that only some of the first stage coefficients come from this common distribution, to take account of settings where some of the instruments are qualitatively different from the others.) The idea is partly that in most cases with many instruments, as for example in the AK study, the instruments arise from interacting a small set of distinct instruments with other covariates. Hence it may be natural to think of the coefficients on these instruments in the reduced form as exchangeable. This notion is captured by modelling the first stage coefficients as independent draws from the same distribution. In addition, this set up parametrizes the many-weak instrument problem in terms of a few parameters: the concern is that the values of both μ_π and σ_π^2 are close to zero.

Assuming also joint normality for (ε_i, η_i), one can derive the likelihood function

$$\mathcal{L}(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2, \Omega).$$

In contrast to the likelihood function in terms of the original parameters $(\beta_0, \beta_1, \pi_0, \pi_1, \Omega)$, this likelihood function depends on a small set of parameters, and a quadratic approximation to its logarithms is more likely to be accurate.

CI discuss some connections between the REQML estimator and LIML and TSLS in the context of this parametric set up. First they show that in large samples, with a large number of instruments, the TSLS estimator corresponds to the restricted maximum likelihood estimator where the variance of the first stage coefficients is fixed at a large number, or $\sigma_\pi^2 = \infty$:

$$\hat{\beta}_{\text{TSLS}} \approx \arg \max_{\beta_0,\beta_1,\pi_0,\mu_\pi} = L(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2 = \infty, \Omega).$$

From a Bayesian perspective, TSLS corresponds approximately to the posterior mode given a flat prior on all the parameters, and thus puts a large amount of prior mass on values of the parameter space where the instruments are jointly powerful.

In the same setting with a large number of instruments, no exogenous covariates, and a known reduced form covariance matrix, the LIML estimator corresponds approximately to the REQML estimator where we fix $\sigma_\pi^2 \cdot (1 \ \beta_1)'\Omega^{-1}(1 \ \beta_1)'$ at a large number. In the special case where we fix $\mu_\pi = 0$ and the random effects specification applies to all isntruments, CI show that the REQML estimator is identical to LIML. However, like the Bekker asymptotics, the REQML calculations suggests that the standard LIML variance is too small: the variance of the REQML estimator is approximately equal to the standard LIML variance times

$$1 + \sigma_\pi^{-2} \cdot \left(\begin{pmatrix} 1 \\ \beta_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \right)^{-1}.$$

This is similar to the Bekker adjustment.

4.3 Choosing Subsets of the Instruments

In an interesting paper Donald and Newey (2001) consider the problem of choosing a subset of an infinite sequence of instruments. They assume the instruments are ordered, so that the choice is the number of instruments to use. Given the set of instruments they consider a variety of estimators including TSLS and LIML. The criterion they focus on is based on an approximation to the expected squared error. This criterion is not feasible because it depends on unknown parameters, but they show that using an estimated version of this leads to approximately the same expected squared error as using the infeasible criterion.

Although in its current form not straightforward to implement, this is a very promising approach that can apply to many related problems such as generalized method of moments settings with many moments.

4.4 Other Estimators

Other estimators have also been investigated in the many weak instrument settings. Hansen, Hausman and Newey (2006), and Hausman, Newey and Woutersen (2007) look at Fuller's estimator, which is modification of LIML that has finite moments. Phillips and Hale (1977) (and later Angrist, Imbens and Krueger, 1999) suggest a jackknive estimator. Hahn, Hausman and Kuersteiner (2004) look at jackknife versions of TSLS.

4.5 Flores' simulations

Many simulations exercises have been carried out for evaluating the performance of testing procedures and point estimators. In general it is difficult to assess the evidence of these experiments. They are rarely tied to actual data sets, and so the choices for parameters, distributions, sample sizes, and number of instruments are typically arbitrary.

In one of the more extensive simulation studies Flores-Lagunes (2007) reports results comparing TSLS, LIML, Fuller, Bias corrected versions of TSLS, LIML and Fuller, a Jacknife version of TSLS (Hahn, Hausman and Kuersteiner, 2004), and the REQML estimator, in settings with 100 and 500 observations, and 5 and 30 instruments for the single endogenous variable. He looks at median bias, median absolute error, inter decile range, coverage rates, and He concludes that "our evidence indicates that the random-effects quasi-maximum likelihood estimator outperforms alternative estimators in terms of median point estimates and coverage rates." Note that Flores-Lagunas does not include LIML with the Bekker standard errors.

References

Anderson, T., and H. Rubin, (1949), "Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations," *Annals of Mathematical Statistics* 21, 570-582-.

Andrews, D., M. Moreira, and J. Stock, (2006), "Optimal Two-sided Invariant Similar Tests for Instrumental Variables Regresion," *Econometrica* 74, 715-752-.

Andrews, D., and J. Stock, (2007), "Inference with Weak Instruments," *Advances in Economics and Econometrics*, Vol III, Blundel,, Newey and Persson (eds.), 122-173.

Angrist, J., G. Imbens, and A. Krueger, (1999), "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics,* 14, 57-67.

Angrist, J., and A. Krueger, (1991), "Does Compulsory Schooling Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics* 106, 979-1014.

Bekker, P., (1994), "Alternative Approximations to the Distribution of Instrumental Variables Estimators," *Econometrica* 62, 657-681.

Bound, J., A. Jaeger, and R. Baker, (1996), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak," *Journal of the American Statistical Association* 90, 443-450.

Card, D., (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica* 69(5), 1127-1160.

Chamberlain, G., and G. Imbens, (2004), "Random Effects Estimators with Many Instrumental Variables," *Econometrica* 72(1), 295-306.

Chao, J., and N. Swanson, (2005), "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica* 73(5), 1673-1692.

Dufour, J.-M., (1997), "Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models," *Econometrica* 65, 1365-1387.

Chioda, L., and M. Janson, (1998), "Optimal Conditional Inference for Instrumental Variables Regression," unpublished manuscript, department of economics, UC Berkeley.

Donald, S., and W. Newey, (2001), "Choosing the Number of Instruments," *Econometrica* 69, 1161-1191.

Flores-Lagunes, A., (2007), "Finite Sample Evidence of IV Estimators Under Weak Instruments," *Journal of Applied Econometrics,* 22, 677-694.

Fuller, W., (1977), "Some Properties of a Modification of the Limited Information Estimator," *Econometrica* 45(), 939-954.

Griliches, Z., (1977), "Estimating the Returns to Schooling – Some Econometric Problems," *Econometrica* 45(1), 1-22.

Hahn, J., and J. Hausman, (2003), "Weak Instruments: Diagnosis and Cures in Empirical Ecnometrics," *American Economic Review, Papers and Proceedings* 93, 118-115.

Hahn, J., J. Hausman, and G. Kuersteiner, (2004), "Estimation with Weak Instruments: Accuracy of Higher Order Bias and MSE Approxmiations," *Econometrics Journal*.

Hansen, C., J. Hausman, and W. Newey, (2006), "Estimation with Many Instrumental Variables," Unpublished Manuscript, Department of Economics, MIT.

Hausman, J., W. Newey, T. Woutersen, J. Chao, and N. Swanson, (2007), "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," Unpublished Manuscript, MIT.

Kleibergen, F., (2002), "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica* 70(5), 1781-1803.

Mikusheva, A., (2007), "Uniform Inferences in Econometrics," Chapter 3, PhD Thesis, Harvard University, Department of Economics.

Moreira, M., (2001), "Tests with Correct Size when Instruments can be Arbitrarily Weak," Unpublished Paper, Department of Economics, Harvard University.

Moreira, M., (2003), "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica* 71(4), 1027-1048.

Philips, P., (1984), "Exact Small Sample Theory in the Simultaneous Equations Model," *Handbook of Econometrics*, (Griliches and Intrilligator, eds), Vol 2, North Holland.

Phillips, G., and C. Hale, (1977), "The Bias of Instrumental Variables Estimators of Simultaneous Equations Systems," *International Economic Review*, 18, 219-228.

Rothenberg, T., (1984), "Approximating the Distributions of Econometric Estimators and Test Statistics," *Handbook of Econometrics*, (Griliches and Intrilligator, eds), Vol 2, Amsterdam, North Holland.

Staiger, D., and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica* 68, 1055-1096.

AEA Lectures **Chicago, IL, January 2012**

Lecture 7b, Wednesday, Jan 7th, 8.00am-9.45am

Generalized Method of Moments and Empirical Likelihood

1. Introduction

Generalized Method of Moments (henceforth GMM) estimation has become an important unifying framework for inference in econometrics in the last twenty years. It can be thought of as nesting almost all the common estimation methods such as maximum likelihood, ordinary least squares, instrumental variables and two–stage–least–squares and nowadays it is an important part of all advanced econometrics text books (Gallant, 1987; Davidson and McKinnon, 1993; Hamilton, 1994; Hayashi, 2000; Mittelhammer, Judge, and Miller, 2000; Ruud, 2000; Wooldridge, 2002). Its formalization by Hansen (1982) centers on the presence of known functions, labelled "moment functions", of observable random variables and unknown parameters that have expectation zero when evaluated at the true parameter values. The method generalizes the "standard" method of moments where expectations of known functions of observable random variables are equal to known functions of the unknown parameters. The "standard" method of moments can thus be thought of as a special case of the general method with the unknown parameters and observed random variables entering additively separable. The GMM approach links nicely to economic theory where orthogonality conditions that can serve as such moment functions often arise from optimizing behavior of agents. For example, if agents make rational predictions with squared error loss, their prediction errors should be orthogonal to elements of the information set. In the GMM framework the unknown parameters are estimated by setting the sample averages of these moment functions, the "estimating equations," as close to zero as possible.

The framework is sufficiently general to deal with the case where the number of moment functions is equal to the number of unknown parameters, the so–called "just–identified case", as well as the case where the number of moments exceeding the number of parameters to be

estimated, the "over–identified case." The latter has special importance in economics where the moment functions often come from the orthogonality of potentially many elements of the information set and prediction errors. In the just-identified case it is typically possible to estimate the parameter by setting the sample average of the moments exactly equal to zero. In the over-identified case this is not feasible. The solution proposed by Hansen (1982) for this case, following similar approaches in linear models such as two– and three–stage–least–squares, is to set a linear combination of the sample average of the moment functions equal to zero, with the dimension of the linear combination equal to the number of unknown parameters. The optimal linear combination of the moments depends on the unknown parameters, and Hansen suggested to employ initial, possibly inefficient, estimates to estimate this optimal linear combination. Chamberlain (1987) showed that this class of estimators achieves the semiparametric efficient bound given the set of moment restrictions. The Chamberlain paper is not only important for its substantive efficiency result, but also as a precursor to the subsequent empirical likelihood literature by the methods employed: Chamberlain uses a discrete approximation to the joint distribution of all the variables to show that the information matrix based variance bound for the discrete parametrization is equal to the variance of the GMM estimator if the discrete approximation is fine enough.

Th empirical likelihood literature developed partly in response to criticisms regarding the small sample properties of the two-step GMM estimator. Researchers found in a number of studies that with the degree of over-identification high, these estimators had substantial biases, and confidence intervals had poor coverage rates. See among others, Altonji and Segal (1996), Burnside and Eichenbaum (1996) , and Pagan and Robertson (1997). These findings are related to the results in the instrumental variables literature that with many or weak instruments two-stage-least squares can perform very badly (e.g., Bekker, 1994; Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997). Simulations, as well as theoretical results, suggest that the new estimators have LIML-like properties and lead to improved large sample properties, at the expense of some computational cost.

2. Examples

First the generic form of the GMM estimation problem in a cross–section context is presented. The parameter vector θ^* is a K dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$. The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector valued function such that $E\left[\psi(Z, \theta^*)\right] = 0$, and $E\left[\psi(Z, \theta)\right] \neq 0$ for all $\theta \in \Theta$ with $\theta \neq \theta^*$. The researcher has available an independent and identically distributed random sample $Z_1, Z_2, \ldots, Z_N$. We are interested in the properties of estimators for θ^* in large samples.

Many, if not most models considered in econometrics fit this framework. Below are some examples, but this list is by no means exhaustive.

I. Maximum Likelihood

If one specifies the conditional distribution of a variable Y given another variable X as $f_{Y|X}(y|x, \theta)$, the score function satisfies these conditions for the moment function:

$$\psi(Y, X, \theta) = \frac{\partial \ln f}{\partial \theta}(Y|X, \theta).$$

By standard likelihood theory the score function has expectation zero only at the true value of the parameter. Interpreting maximum likelihood estimators as generalized method of moments estimators suggests a way of deriving the covariance matrix under misspecification (e.g., White, 1982), as well as an interpretation of the estimand in that case.

II. Linear Instrumental Variables

Suppose one has a linear model

$$Y = X'\theta^* + \varepsilon,$$

with a vector of instruments Z. In that case the moment function is

$$\psi(Y, X, Z, \theta) = Z' \cdot (Y - X'\theta).$$

The validity of Z as an instrument, together with a rank condition implies that θ^* is the unique solution to $E[\psi(Y, X, Z, \theta)] = 0$. This is a case where the fact that the methods allow for more moments than unknown parameters is of great importance as often instruments are

independent of structural error terms, implying that any function of the basic instruments is orthogonal to the errors.

III. A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel data models studied in Keane and Runkle (1992), Chamberlain (1992), and Blundell and Bond (1998). This specific model has previously been studied by Bond, Bowsher, and Windmeijer (2001). One can construct moment functions by differencing and using lags as instruments, as in Arellano and Bond (1991), and Ahn and Schmidt, (1995):

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3, \ldots, T$, leading to a total of $(T-1) \cdot (T-2)/2$ moments, with only a single parameter. One would typically expect that the long lags do not necessarily contain much information, but they are often used to improve efficiency. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

Despite the different nature of the two sets of moment functions, which makes them potentially very useful in the case that the autoregressive parameter is close to unity, they can all be combined in the GMM framework.

3. Two-step GMM Estimation

3.1 Estimation and Inference

In the just-identified case where M, the dimension of ψ, and K, the dimension of θ are identical, one can generally estimate θ^* by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \hat{\theta}_{\rm gmm}). \tag{1}$$

If the sample average is replaced by the expectation, the unique solution is equal to θ^*, and under regularity conditions (e.g., Hansen, 1982, Newey and McFadden, 1994), solutions to (1) will be unique in large samples and consistent for θ^*. If $M > K$ the situation is more complicated as in general there will be no solution to (1).

Hansen's (1982) solution was to generalize the optimization problem to the minimization of the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i,\theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N} \psi(z_i,\theta)\right], \tag{2}$$

for some positive definite $M \times M$ symmetric matrix C. Under the regularity conditions given in Hansen (1982) and Newey and McFadden (1994), the minimand $\hat{\theta}_{\rm gmm}$ of (2) has the following large sample properties:

$$\hat{\theta}_{\rm gmm} \xrightarrow{p} \theta^*,$$

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1}\Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}),$$

where

$$\Delta = \mathbb{E}\left[\psi(Z_i,\theta^*)\psi(Z_i,\theta^*)'\right] \qquad \text{and} \quad \Gamma = \mathbb{E}\left[\frac{\partial}{\partial \theta'}\psi(Z_i,\theta^*)\right].$$

In the just–identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial, as $\hat{\theta}_{\rm gmm}$ will, at least in large samples, be equal to the value of θ that sets the average moments exactly equal to zero. In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible and the asymptotic covariance matrix reduces to $(\Gamma'\Delta^{-1}\Gamma)^{-1}$, irrespective of the choice of C. In the overidentified case with $M > K$, however, the choice of the weight matrix C is important. The optimal choice for C in terms of minimizing the asymptotic variance is in this case

the inverse of the covariance of the moments, Δ^{-1}. Using the optimal weight matrix, the asymptotic distribution is

$$\sqrt{N}(\hat{\theta}_{\rm gmm} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1}). \tag{3}$$

This estimator is generally not feasible because typically Δ^{-1} is not known to the researcher. The feasible solution proposed by Hansen (1982) is to obtain an initial consistent, but generally inefficient, estimate of θ^* by minimizing $Q_{C,N}(\theta)$ using an arbitrary positive definite $M \times M$ matrix C, e.g., the identity matrix of dimension M. Given this initial estimate, $\tilde{\theta}$, one can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})'\right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\hat{\Delta}^{-1},N}(\theta)$. The resulting estimator $\hat{\theta}_{\rm gmm}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Hansen (1982) also suggested a specification test for this model. If the number of moments exceeds the number of free parameters, not all average moments can be set equal to zero, and their deviation from zero forms the basis of Hansen's test, similar to tests developed by Sargan (1958). See also Newey (1985a, 1985b). Formally, the test statistic is

$$T = Q_{\hat{\Delta},N}(\hat{\theta}_{\rm gmm}).$$

Under the null hypothesis that all moments have expectation equal to zero at the true value of the parameter, θ^*, the distribution of the test statistic converges to a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions, $M - K$.

One can also interpret the two–step estimator for over–identified GMM models as a just–identified GMM estimator with an augmented parameter vector (e.g., Newey and McFadden, 1994; Chamberlain and Imbens, 1995). Define the following moment function:

$$h(x,\delta) = h(x,\theta,\Gamma,\Delta,\beta,\Lambda) = \begin{pmatrix} \Lambda - \frac{\partial\psi}{\partial\theta'}(x,\beta) \\ \Lambda' C\psi(x,\beta) \\ \Delta - \psi(x,\beta)\psi(x,\beta)' \\ \Gamma - \frac{\partial\psi}{\partial\theta'}(x,\theta) \\ \Gamma'\Delta^{-1}\psi(x,\theta) \end{pmatrix}. \tag{4}$$

Because the dimension of the moment function $h(\cdot)$, $M\times K+K+(M+1)\times M/2+M\times K+K = (M+1)\times(2K+M/2)$, is equal to the combined dimensions of its parameter arguments, the estimator for $\delta=(\theta,\Gamma,\Delta,\beta,\Lambda)$ obtained by setting the sample average of $h(\cdot)$ equal to zero is a just–identified GMM estimator. The first two components of $h(x,\delta)$ depend only on β and Λ, and have the same dimension as these parameters. Hence β^* and Λ^* are implicitly defined by the equations

$$E\left[\left(\begin{array}{c}\Lambda-\frac{\partial\psi}{\partial\theta'}(X,\beta)\\ \Lambda' C\psi(X,\beta)\end{array}\right)\right]=0.$$

Given β^* and Λ^*, Δ^* is defined through the third component of $h(x,\delta)$, and given β^*, Λ^* and Δ^* the final parameters θ^* and Γ^* are defined through the last two moment functions.

This interpretation of the over-identified two-step GMM estimator as a just-identified GMM estimator in an augmented model is interesting because it also emphasizes that results for just–identified GMM estimators such as the validity of the bootstrap can directly be translated into results for over–identified GMM estimators. In another example, using the standard approach to finding the large sample covariance matrix for just–identified GMM estimators one can use the just-identified representation to find the covariance matrix for the over–identified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial\delta}(X,\delta^*)\right]\right)^{-1}E[h(Z,\delta^*)h(Z,\delta^*)']\left(E\left[\frac{\partial h}{\partial\delta}(Z,\delta^*)\right]\right)^{-1},$$

estimated by averaging at the estimated values. This is the GMM analogue of the White (1982) covariance matrix for the maximum likelihood estimator under misspecification.

3.2 Efficiency

Chamberlain (1987) demonstrated that Hansen's (1982) estimator is efficient, not just in the class of estimators based on minimizing the quadratic form $Q_{N,C}(\theta)$, but in the larger class of semiparametric estimators exploiting the full set of moment conditions. What is particularly interesting about this argument is the relation to the subsequent empirical likelihood literature. Many semiparametric efficiency bound arguments (e.g., Newey, 1991; Hahn, 1994)

implicitly build fully parametric models that include the semiparametric one and then search for the least favorable parametrization. Chamberlain's argument is qualitatively different. He proposes a specific parametric model that can be made arbitrarily flexible, and thus arbitrarily close to the model that generated the data, but does not typically include that model. The advantage of the model Chamberlain proposes is that it is in some cases very convenient to work with in the sense that its variance bound can be calculated in a straightforward manner. The specific model assumes that the data are discrete with finite support $\{\lambda_1, \ldots, \lambda_L\}$, and unknown probabilities $\pi_1, \ldots, \pi_L$. The parameters of interest are then implicitly defined as functions of these points of support and probabilities. With only the probabilities unknown, the variance bound on the parameters of the approximating model are conceptually straightforward to calculate. It then sufficies to translate that into a variance bound on the parameters of interest. If the original model is over-identified, one has restrictions on the probabilities. These are again easy to evaluate in terms of their effect on the variance bound.

Given the discrete model it is straightforward to obtain the variance bound for the probabilities, and thus for any function of them. The remarkable point is that one can rewrite these bounds in a way that does not involve the support points. This variance turns out to be identical to the variance of the two-step GMM estimator, thus proving its efficiency.

4. Empirical Likelihood

4.1 Background

To focus ideas, consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. If we wish to estimate the common distribution of these random variables, the natural choice is the empirical distribution, that puts weight $1/N$ on each of the N sample points. However, in a GMM setting this is not necessarily an appropriate estimate. Suppose the moment function is

$$\psi(z, \theta) = z,$$

implying that the expected value of Z is zero. Note that in this simple example this moment

function does not depend on any unknown parameter. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\hat{F}_{emp}}[Z] = \sum z_i/N \neq 0$. The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\hat{F}$ does satisfy the restriction. In other words, the approach is to look for the distribution function closest to $\hat{F}_{\rm emp}$, within the set of distribution functions satisfying $E_F[Z] = 0$. Empirical likelihood provides an operationalization of the concept of closeness here. The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i,$$

for $0 \leq \pi_i \leq 1$, $\sum_{i=1}^{N} \pi_i = 1$. This is not a likelihood function in the standard sense, and thus does not have all the properties of likelihood functions. The empirical likelihood estimator for the distribution function is

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i z_i = 0.$$

Without the second restriction the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied. In this example the solution for the Lagrange multiplier is the solution to the equation

$$\sum_{i=1}^{N} \frac{z_i}{1 + t \cdot z_i} = 0,$$

and the solution for π_i is:

$$\hat{\pi}_i = 1/(1 + t \cdot z_i).$$

More generally, in the over-identified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta,\pi} \sum_{i=1}^{N} \ln \pi, \qquad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0. \tag{5}$$

Qin and Lawless (1994) and Imbens (1997) show that this estimator is equivalent, to order $O_p(N^{-1/2})$, to the two-step GMM estimator. This simple discussion illustrates that for some,

and in fact many, purposes the empirical likelihood has the same properties as a parametric likelihood function. This idea, first proposed by Owen (1988), turns out to be very powerful with many applications. Owen (1988) shows how one can construct confidence intervals and hypothesis tests based on this notion.

Related ideas have shown up in a number of places. Cosslett's (1981) work on choice-based sampling can be interpreted as maximizing a likelihood function that is the product of a parametric part coming from the specification of the conditional choice probabilities, and an empirical likelihood function coming from the distribution of the covariates. See Imbens (1992) for a connection between Cosslett's work and two-step GMM estimation. As mentioned before, Chamberlain's (1987) efficiency proof essentially consists of calculating the distribution of the empirical likelihood estimator and showing its equivalence to the distribution of the two-step GMM estimator. See Back and Brown (1990) and Kitamura and Stutzer (1997) for a discussion of the dependent case.

4.2 Cressie-Read Discrepancy Statistics and Generalized Empirical Likelihood

In this section we consider a generalization of the empirical likelihood estimators based on modifications of the objective function. Corcoran (1998) (see also Imbens, Spady and Johnson, 1998), focus on the Cressie-Read discrepancy statistic, for fixed λ, as a function of two vectors p and q of dimension N (Cressie and Read 1984):

$$I_\lambda(p,q) = \frac{1}{\lambda\cdot(1+\lambda)}\sum_{i=1}^N p_i\left[\left(\frac{p_i}{q_i}\right)^\lambda - 1\right].$$

The Cressie-Read minimum discrepancy estimators are based on minimizing this difference between the empirical distribution, that is, the N-dimensional vector with all elements equal to $1/N$, and the estimated probabilities, subject to all the restrictions being satisfied.

$$\min_{\pi,\theta} I_\lambda(\iota/N,\pi) \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i\cdot\psi(z_i,\theta) = 0.$$

If there are no binding restrictions, because the dimension of $\psi(\cdot)$ and θ agree (the just-identified case), the solution for π is the empirical distribution it self, and $\pi_i = 1/N$. More

generally, if there are over-identifying restrictions, there is no solution for θ to $\sum_i \psi(z_i,\theta)/N = 0$, and so the solution for π_i is as close as possible to $1/N$ in a way that ensures there is an exact solution to $\sum_i \pi_i\psi(z_i,\theta) = 0$. The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Three special cases of this class have received most attention. First, the empirical likelihood estimator itself, which can be interpreted as the case with $\lambda \to 0$. This has the nice interpretation that it is the exact maximum likelihood estimator if Z has a discrete distribution. It does not rely on the discreteness for its general properties, but this interpretation does suggest that it may have attractive large sample properties.

The second case is the exponential tilting estimator with $\lambda \to -1$ (Imbens, Spady and Johnson, 1998), whose objective function is equal to the empirical likelihood objective funtion with the role of π and ι/N reversed. It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^N \pi_i \ln \pi_i \qquad \text{subject to } \sum_{i=1}^N \pi_i = 1, \quad \text{and } \sum_{i=1}^N \pi_i\psi(z_i,\theta) = 0.$$

Third, the case with $\lambda = -2$. This case was originally proposed by Hansen, Heaton and Yaron (1996) as the solution to

$$\min_\theta \frac{1}{N}\left[\sum_{i=1}^N \psi(z_i,\theta)\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\right]^{-1} \cdot \left[\sum_{i=1}^N \psi(z_i,\theta)\right],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments. Hansen, Heaton and Yaron (1996) labeled this the continuously updating estimator. Newey and Smith (2004) pointed out that this estimator fits in the Cressie-Read class.

Smith (1997) considers a more general class of estimators, which he labels generalized empirical likelihood estimators, starting from a different perspective. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$, $g'(0) = 1$, consider the saddle point problem

$$\max_\theta \min_t \sum_{i=1}^N g(t'\psi(z_i,\theta)).$$

This representation is more attractive from a computational perspective, as it reduces the dimension of the optimization problem to $M + K$ rather than a constrained optimization problem of dimension $K + N$ with $M + 1$ restrictions. There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie-Read representation. Newey and Smith (2004) how to choose $g(\cdot)$ for a given λ so that the corresponding GEL and Cressie-Read estimators agree.

In general the differences between the estimators within this class is relatively small compared to the differences between them and the two-step GMM estimators. In practice the choice between them is largely driven by computational issues, which will be discussed in more detail in Section 5. The empirical likelihood estimator does have the advantage of its exact likelihood interpretation and the resulting optimality properties for its bias-corrected version (Newey and Smith, 2004). On the other hand, Imbens, Spady and Johnson (1998) argue in favor of the exponential tilting estimator as its influence function stays bounded where as denominator in the probabilities in the empirical likelihood estimator can get large. In simulations researcher have encountered more convergence problems with the continuously updating estimator (e.g., Hansen, Heaton and Yaron, 1996; Imbens, Johnson and Spady, 1998).

4.3 Testing

Associated with the empirical likelihood estimators are three tests for over-identiyfing restrictions, similar to the classical trinity of tests, the likelihood ratio, the Wald, and the Lagrange multiplier tests. Here we briefly review the implementation of the three tests in the empirical likelihood context. The leading terms of all three tests are identical to that of the test developed by Hansen (1982) based on the quadratic form in the average moments.

The first test is based on the value of the empirical likelihood function. The test statistic compares the value of the empirical likelihood function at the restricted estimates, the $\hat{\pi}_i$

with that at the unrestricted values, $\pi_i = 1/N$:

$$LR = 2 \cdot (L(\iota/N) - L(\hat{\pi})), \qquad \text{where} \quad L(\pi) = \sum_{i=1}^{N} \ln \pi_i.$$

As in the parametric case, the difference between the restricted and unrestricted likelihood function is multiplied by two to obtain, under regularity conditions, e.g., Newey and Smith (2004), a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions for the test statistic under the null hypothesis.

The second test, similar to Wald tests, is based on the difference between the average moments and their probability limit under the null hypothesis, zero. As in the standard GMM test for overidentifying restrictions (Hansen, 1982), the average moments are weighted by the inverse of their covariance matrix:

$$Wald = \frac{1}{N} \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right]' \hat{\Delta}^{-1} \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right],$$

where $\hat{\Delta}$ is an estimate of the covariance matrix

$$\Delta = E[\psi(Z, \theta^*)\psi(Z, \theta)'],$$

typically based on a sample average at some consistent estimator for θ^*:

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^{N} \psi(z_i, \hat{\theta})\psi(z_i, \hat{\theta})',$$

or sometimes a fully efficient estimator for the covariance matrix,

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\pi}_i \psi(z_i, \hat{\theta})\psi(z_i, \hat{\theta})',$$

The standard GMM test uses an initial estimate of θ^* in the estimation of Δ, but with the empirical likelihood estimators it is more natural to substitute the empirical likelihood estimator itself. The precise properties of the estimator for Δ do not affect the large sample properties of the test, and like the likelihood ratio test, the test statistic has in large samples a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions.

The third test is based on the Lagrange multipliers t. In large samples their variance is

$$V_t = \Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}.$$

This matrix is singular, with rank equal to $M - K$. One option is therefore to compare the Lagrange multipliers to zero using a generalized inverse of their covariance matrix:

$$LM_1 = t' \left(\Delta^{-1} - \Delta^{-1}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\right)^{-g} t.$$

This is not very attractive, as it requires the choice of a generalized inverse. An alternative is to use the inverse of Δ^{-1} itself, leading to the test statistic

$$LM_2 = t'\Delta t.$$

Because

$$\sqrt{N} \cdot t = V_t \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(z_i, \theta^*) + o_p(1),$$

and $V_t \Delta V_t = V_t V_t^{-g} V_t = V_t$, it follows that

$$LM_2 = LM_1 + o_p(1).$$

Imbens, Johnson and Spady (1998) find in their simulations that tests based on LM_2 perform better than those based on LM_1. In large samples both have a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions. Again we can use this test with any efficient estimator for t, and with the Lagrange multipliers based on any of the discrepancy measures.

Imbens, Spady and Johnson (1998), and Bond, Bowsher and Windmeijer (2001) investigate through simulations the small sample properties of various of these tests. It appears that the Lagrange multiplier tests are often more attractive than the tests based on the average moments, although there is so far only limited evidence in specific models. One can use the same ideas for constructing confidence intervals that do not directly use the normal approximation to the sampling distribution of the estimator. See for discussions Smith (1998) and Imbens and Spady (2002).

6. Computational Issues

The two-step GMM estimator requires two minimizations over a K-dimensional space. The empirical likelihood estimator in its original likelihood form (5) requires maximization over a space of dimension K (for the parameter θ) plus N (for the N probabilities), subject to $M+1$ restrictions (on the M moments and the adding up restriction for the probabilities). This is in general a much more formidable computational problem than two optimizations in a K-dimensional space. A number of approaches have been attempted to simplify this problem. Here we disuss three of them in the context of the exponential tilting estimator, although most of them directly carry over to other members of the Cressie-Read or GEL classes.

6.1 Solving the First Order Conditions

The first approach we discuss is focuses on the first order conditions and then concentrates out the probabilities π. This reduces the problem to one of dimension $K+M$, K for the parameters of interest and M for the Lagrange multipliers for the restrictions, which is is clearly a huge improvement, as the dimension of the problem no longer increases with the sample size. Let μ and t be the Lagrange multipliers for the restrictions $\sum \pi_i = 1$ and $\sum \pi_i \psi(z_i,\theta) = 0$. The first order conditions for the π's and θ and the Lagrange multipliers are

$$
\begin{aligned}
0 &= \ln \pi_i - 1 - \mu + t'\psi(z_i,\theta), \\
0 &= \sum_{i=1}^N \pi_i \frac{\partial \psi}{\partial \theta'}(z_i,\theta), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \exp\left(t'\psi(z_i,\theta)\right), \\
0 &= \exp(\mu - 1) \sum_{i=1}^N \psi(z_i,\theta) \cdot \exp\left(t'\psi(z_i,\theta)\right).
\end{aligned}
$$

The solution for π is

$$
\pi_i = \exp(\mu - 1 + t'\psi(z_i,\theta)).
$$

To determine the Lagrange multipliers t and the parameter of interest θ we only need π_i up to a constant of proportionality, so we can solve

$$0 = \sum_{i=1}^{N} \psi(z_i, \theta) \exp(t'\psi(z_i, \theta)), \tag{6}$$

and

$$0 = \sum_{i=1}^{N} t' \frac{\partial \psi}{\partial \theta}(z_i, \theta) \exp(t'\psi(z_i, \theta)) \tag{7}$$

Solving the system of equations (6) and (7) is not straightforward. Because the probability limit of the solution for t is zero, the derivative with respect to θ of both first order conditions converges zero. Hence the matrix of derivatives of the first order conditions converges to a singular matrix. As a result standard approaches to solving systems of equations can behave erratically, and this approach to calculating $\hat{\theta}$ has been found to have poor operating characteristics.

6.2 Penalty Function Approaches

Imbens, Spady and Johnson (1998) characterize the solution for θ and t as

$$\max_{\theta,t} K(t, \theta) \qquad \text{subject to } K_t(t, \theta) = 0, \tag{8}$$

where $K(t, \theta)$ is the empirical analogue of the cumulant generating function:

$$K(t, \theta) = \ln \left[\frac{1}{N} \sum_{i=1}^{N} \exp(t'\psi(z_i, \theta) \right].$$

They suggest solving this optimization problem by maximizing the unconstrained objective function with a penalty term that consists of a quadratic form in the restriction:

$$\max_{\theta,t} K(t, \theta) - 0.5 \cdot A \cdot K_t(t, \theta)' W^{-1} K_t(t, \theta), \tag{9}$$

for some positive definite $M \times M$ matrix W, and a positive constant A. The first order conditions for this problem are

$$0 = K_\theta(t, \theta) - A \cdot K_{t\theta}(t, \theta) W^{-1} K_t(t, \theta),$$

$$0 = K_t(t,\theta) - A \cdot K_{tt}(t,\theta)W^{-1}K_t(t,\theta).$$

For A large enough the solution to this unconstrained maximization problem is identical to the solution to the constrained maximization problem (8). This follows from the fact that the constraint is in fact the first order condition for $K(t,\theta)$. Thus, in contrast to many penalty function approaches, one does not have to let the penalty term go to infinity to obtain the solution to the constrained optimization problem, one only needs to let the penalty term increase sufficiently to make the problem locally convex. Imbens, Spady and Johnson (1998) suggest choosing

$$W = K_{tt}(t,\theta) + K_t(t,\theta)K_t(t,\theta)',$$

for some initial values for t and θ as the weight matrix, and report that estimates are generally not sensitive to the choices of t and θ.

6.3 Concentrating out the Lagrange Multipliers

Mittelhammer, Judge and Schoenberg (2001) suggest concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. As shown above, concentrating out the probabilities π_i can be done analytically. Although it is not in general possible to solve for the Lagrange multipliers t analytically, other than in the continuously updating case, for given θ it is easy to numerically solve for t. This involves solving, in the exponential tilting case,

$$\min_t \sum_{i=1}^N \exp(t'\psi(z_i,\theta)).$$

This function is strictly convex as a function of t, with the easy to calculate first and second derivatives equal to

$$\sum_{i=1}^N \psi(z_i,\theta)\exp(t'\psi(z_i,\theta)),$$

and

$$\sum_{i=1}^N \psi(z_i,\theta)\psi(z_i,\theta)'\exp(t'\psi(z_i,\theta)),$$

respectively. Therefore concentrating out the Lagrange multipliers is computationally fast using a Newton-Raphson algorithm. The resulting function $t(\theta)$ has derivatives with respect to θ equal to:

$$\frac{\partial t}{\partial \theta'}(\theta) = -\left(\frac{1}{N}\sum_{i=1}^{N}\psi(z_i,\theta)\psi(z_i,\theta)'\exp(t(\theta)'\psi(z_i,\theta)\right)^{-1}$$
$$\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi}{\partial\theta'}(z_i,\theta)\exp(t(\theta)'\psi(z_i,\theta)+\psi(z_i,\theta)t(\theta)'\frac{\partial\psi}{\partial\theta'}(z_i,\theta)\exp(t(\theta)'\psi(z_i,\theta))\right)$$

After solving for $t(\theta)$, one can solve

$$\max_{\theta}\sum_{i=1}^{N}\exp(t(\theta)'\psi(z_i,\theta)). \tag{10}$$

Mittelhammer, Judge, and Schoenberg (2001) use methods that do not require first derivatives to solve (10). This is not essential. Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ. In general these are straightforward to calculate and likely to improve the performance of the algorithm.

In this method in the end the researcher only has to solve one optimization in a K-dimensional space, with the provision that for each evaluation of the objective function one needs to numerically evaluate the function $t(\theta)$ by solving a convex maximization problem. The latter is fast, especially in the exponential tilting case, so that although the resulting optimization problem is arguably still more difficult than the standard two-step GMM problem, in practice it is not much slower. In the simulations below I use this method for calculating the estimates. After concentrating out the Lagrange multipliers using a Newton-Rahpson algorithm that uses both first and second derivatives, I use a Davidon-Fletcher-Powell algorithm to maximize over θ, using analytic first derivatives. Given a direction I used a line search algoritm based on repeated quadratic approximations.

7. A Dynamic Panel Data Model

To get a sense of the finite sample properties of the empirical likelihood estimators we compare some of the GMM methods in the context of the panel data model briefly discussed

in Section 2, using some simulation results from Imbens. The model is

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T. This is a stylized version of the type of panel data models extensively studied in the literature. Bond, Bowsher and Windmeijer (2001) study this and similar models to evaluate the performance of test statistics based on different GMM and gel estimators. We use the moments

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3, \ldots, T$, leading to a total of $(T-1) \cdot (T-2)/2$ moments. In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

It is important to note, given the results discussed in Section 4, that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. As a result there is potentially a substantial difference between the different estimators, especially when the degree of overidentification is high.

We report some simulations for a data generating process with parameter values estimated on data from Abowd and Card (1989) taken from the PSID. See also Card (1994). This data set contains earnings data for 1434 individuals for 11 years. The individuals are selected on having positive earnings in each of the eleven years, and we model their earnings in logarithms. We focus on estimation of the autoregressive coefficient θ.

We then generate artificial data sets to investigate the repeated sampling properties of these estimators. Two questions are of most interest. First, how do the median bias and

median-absolute-error deteriorate as a function of the degree of over-identification? Here, unlike in the theoretical discussion in Section 4, the additional moments, as we increase the number of years in the panel, do contain information, so they may in fact increase precision, but at the same time one would expect based on the theoretical calculations that the accuracy of the asymptotic approximations for a fixed sample size deteriorates with the number of years. Second, we are interested in the performance of the confidence intervals for the parameter of interest. In two-stage-least-squares settings it was found that with many weak instruments the performance of standard confidence intervals varied widely between liml and two-stage-least-squares estimators. Given the analogy drawn by Hansen, Heaton and Yaron (1996) between the continuously updating estimator and liml, the question arises how the confidence intervals differ between two-step GMM and the various Cressie-Read and GEL estimators.

Using the Abowd-Card data we estimate θ and the variance of the fixed effect and the idiosyncratic error term. The latter two are estimated to be around 0.3. We then consider data generating processes where the individual effect η_i has mean zero and standard deviation equal to 0.3, and the error term has mean zero and standard deviation 0.3. We $\theta = 0.9$ in the simulations. This is larger than the value in estimated from the Abowd-Card data. We compare the standard Two-Step GMM estimator and the Exponential Tilting Estimator. Table 1 contains the results. With the high autoregressive coefficient, $\theta = 0.9$, the two-step GMM estimator has substantial bias and poor coverage rates.The exponential tilting estimator does much better with the high autoregressive coefficient. The bias is small, on the order of 10% of the standard error, and the coverage rate is much closer to the nominal one.

References

Abowd, J. and D. Card, (1989), "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57 (2), 441-445.

Ahn, S., and P. Schmidt, (1995), "Efficient Estimation of Models for Dynamic Panel Data", *Journal of Econometrics*, 68, 5-28.

Altonji, J., and L. Segal, (1996), "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, Vol 14, No. 3, 353–366.

Back, K., and D. Brown, (1990), "Estimating Distributions from Moment Restrictions", working paper, Graduate School of Business, Indiana University.

Bekker, P., (1994), "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 62, 657-681.

Bond, S., C. Bowsher, and F. Windmeijer, (2001), "Criterion-based Inference for GMM in Linear Dynamic Panel Data Models", IFS, London.

Bound, J., D. Jaeger, and R. Baker, (1995), "Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak", forthcoming, em Journal of the American Statistical Association.

Burnside, C., and M., Eichenbaum, (1996), "Small Sample Properties of Generalized Method of Moments Based Wald Tests", *Journal of Business and Economic Statistics*, Vol. 14, 294-308.

Card, D., (1994) "Intertemporal Labour Supply: an Assessment", in: *Advances in Econometrics*, Simms (ed), Cambridge University Press.

Chamberlain, G., (1987), "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions", *Journal of Econometrics*, vol. 34, 305–334, 1987

Corcoran, S., (1998), "Bartlett Adjustment of Empirical Discrepancy Statistics", *Biometrika*.

Cosslett, S. R., (1981), "Maximum Likelihood Estimation for Choice–based Samples", *Econometrica*, vol. 49, 1289–1316,

Cressie, N., and T. Read, (1984), "Multinomial Goodness-of-Fit Tests", *Journal of the Royal Statistical Society*, Series B, 46, 440-464.

Hall, A., (2005), *Generalized Method of Moments*, Oxford University Press.

Hansen, L–P., (1982), "Large Sample Properties of Generalized Method of Moment Estimators", *Econometrica*, vol. 50, 1029–1054.

Hansen, L.–P., J. Heaton, and A. Yaron, (1996), "Finite Sample Properties of Some Alternative GMM Estimators", *Journal of Business and Economic Statistics*, Vol 14, No. 3, 262–280.

Imbens, G. W., (1992), "Generalized Method of Moments and Empirical Likelihood," *Journal of Business and Economic Statistics*, vol. 60.

Imbens, G. W., R. H. Spady, and P. Johnson, (1998), "Information Theoretic Approaches to Inference in Moment Condition Models", *Econometrica*.

Imbens, G., and R. Spady, (2002), "Confidence Intervals in Generalized Method of Moments Models," *Journal of Econometrics*, 107, 87-98.

Imbens, G., and R. Spady, (2005), "The Performance of Empirical Likelihood and Its Generalizations," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Kitamura, Y., and M. Stutzer, (1997), "An Information-theoretic Alternative to Generalized Method of Moments Estimation", *Econometrica*, Vol. 65, 861-874.

Mittelhammer, R., G. Judge, and R. Schoenberg, (2005), "Empirical Evidence Concerning the Finite Sample Performance of EL-Type Structural Equation Estimation and Inference Methods," in *Identification and Inference for Econometric Models, Essays in Honor of Thomas Rothenberg*, Andrews and Stock (eds).

Mittelhammer, R., G. Judge, and D. Miller, (2000), *Econometric Foundations*, Cambridge University Press, Cambridge.

Newey, W., (1985), "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, vol. 29, 229–56.

Newey, W., and D. McFadden, (1994) "Estimation in Large Samples", in: McFadden and Engle (Eds.), *The Handbook of Econometrics*, Vol. 4.

Newey, W., and R. Smith, (2004), "Higher Order Properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 573-595.

Owen, A., (1988), "Empirical Likelihood Ratios Confidence Intervals for a Single Functional", *Biometrika*, 75, 237-249.

Owen, A., (2001), *Empirical Likelihood*, Chapman and Hall, London.

Pagan, A., and J. Robertson, (1997), "GMM and its Problems", unpublished manuscript, Australian National University.

Qin, and J. Lawless, (1994), "Generalized Estimating Equations", *Annals of Stat.*

Smith, R., (1997), "Alternative Semiparametric Likelihood Approaches to Generalized Method of Moments Estimation", *Economic Journal*, 107, 503-19.

Staiger, D., and J. Stock, (1997), "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557-586.

White, H., (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, vol. 50, 1–25.

Wooldridge, J., (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples", *Econometrica* 67, No. 6, 1385-1406.

Table 1: SIMULATIONS, $\theta = 0.9$

	Number of time periods								
	3	4	5	6	7	8	9	10	11
Two-Step GMM									
median bias	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.03	0.08	0.03	0.11	0.08	0.13	0.11
median absolute error	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.88	0.85	0.82	0.80	0.80	0.79	0.78	0.79	0.76
covarage rate 95% ci	0.92	0.91	0.89	0.87	0.85	0.86	0.86	0.88	0.84
Exponential Tilting									
median bias	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	0.00	0.00
relative median bias	0.04	0.09	0.02	-0.00	0.01	0.01	-0.02	0.08	0.13
median absolute error	0.05	0.03	0.03	0.02	0.02	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.87	0.86	0.84	0.86	0.88	0.86	0.87	0.88	0.87
covarage rate 95% ci	0.91	0.90	0.90	0.91	0.93	0.92	0.91	0.93	0.93

The relative median bias reports the bias divided by the large sample standard error. All results based on 10,000 replications.

"Cross-Section Econometrics"

Lecture 7

Many and Weak Instruments, GMM and EL

Guido Imbens

AEA Lectures, Chicago, January 2012

Part I: Weak and Many Instruments

1. Introduction
2. Motivation
3. Weak Instruments
4. Many (Weak) Instruments

1. Introduction

Standard normal asymptotic approximation to sampling distribution of IV, TSLS, and LIML estimators relies on non-zero correlation between instruments and endogenous regressors.

If correlation is close to zero, these approximations are not accurate, even in fairly large samples.

In the just identified case TSLS/LIML confidence intervals will still be fairly wide in most cases, even if not valid, unless degree of endogeneity is extremely high. If this is a concern, alternative confidence intervals, based on Anderson-Rubin statistic, can be used: these are valid uniformly. No better estimators available.

In the case with large degree of overidentification TSLS has poor properties: considerable bias towards OLS, and substantial underestimation of standard errors.

LIML is much better in terms of bias, but its standard error is not correct. A simple multiplicative adjustment to conventional LIML standard errors based on Bekker many-instrument asymptotics or random effects likelihood works well.

Overall: use LIML, with Bekker-adjusted standard errors.

2.A Motivation : Angrist-Krueger

AK were interested in estimating the returns to years of education. Their basic specification is:

$$Y_i = \alpha + \beta \cdot E_i + \varepsilon_i,$$

where Y_i is log (yearly) earnings and E_i is years of education.

In an attempt to address the endogeneity problem AK exploit variation in schooling levels that arise from differential impacts of compulsory schooling laws by quarter of birth and use quarter of birth as an instrument. This leads to IV estimate (using only 1st and 4th quarter data):

$$\hat{\beta} = \frac{\overline{Y}_4 - \overline{Y}_1}{\overline{E}_4 - \overline{E}_1} = 0.089 \quad (0.011)$$

2.B AK with Many Instruments

AK also present estimates based on additional instruments. They take the basic 3 qob dummies and interact them with 50 state and 9 year of birth dummies.

Here (following Chamberlain and Imbens) we interact the single binary instrument with state times year of birth dummies to get 500 instruments. Also including the state times year of birth dummies as exogenous covariates leads to the following model:

$$Y_i = X_i'\beta + \varepsilon_i, \qquad \mathbb{E}[Z_i \cdot \varepsilon_i] = 0,$$

where X_i is the 501-dimensional vector with the 500 state/year dummies and years of education, and Z_i is the vector with 500 state/year dummies and the 500 state/year dummies multiplying the indicator for the fourth quarter of birth.

The TSLS estimator for β is

$$\hat{\beta}_{\text{TSLS}} = 0.073 \quad (0.008)$$

suggesting the extra instruments improve the standard errors a little bit.

However, LIML estimator tells a somewhat different story,

$$\hat{\beta}_{\text{LIML}} = 0.095 \quad (0.017)$$

with an increase in the standard error.

1.C Bound-Jaeger-Baker Critique

BJB suggest that despite the large (census) samples used by AK asymptotic normal approximations may be very poor because the instruments are only very weakly correlated with the endogenous regressor.

The most striking evidence for this is based on the following calculation. Take the AK data and re-calculate their estimates after replacing the actual quarter of birth dummies by random indicators with the same marginal distribution.

In principle this means that the standard (gaussian) large sample approximations for TSLS and LIML are invalid since they rely on non-zero correlations between the instruments and the endogenous regressor.

	Single Instr		500 Instruments			
			TSLS		LIML	
Real QOB	0.089	(0.011)	0.073	(0.008)	0.095	(0.017) [0.037]
Random QOB	0.181	(0.193)	0.059	(0.009)	-0.134	(0.065) [0.251]

With many random instruments the results are troubling. Although the instrument contains no information, the results suggest that the instruments can be used to infer precisely the returns to education.

1.D Simulations with a Single Instrument

10,000 artificial data sets, all of size 160,000, designed to mimic the AK data. In each of these data sets half the units have quarter of birth (denoted by Q_i) equal to 0 and 1 respectively.

$$\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.446 & \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} \\ \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} & 10.071 \end{pmatrix} \right).$$

The correlation between the reduced form residuals in the AK data is $\rho = 0.318$.

$$E_i = 12.688 + 0.151 \cdot Q_i + \eta_i,$$

$$Y_i = 5.892 + 0.014 \cdot Q_i + \nu_i.$$

Now we calculate the IV estimator and its standard error, using either the actual qob variable or a random qob variable as the instrument.

We are interested in the size of tests of the null that coefficient on years of education is equal to 0.089(= 0.014/0.151).

We base the test on the t-statistic. Thus we reject the null if the ratio of the point estimate minus 0.089 and the standard error is greater than 1.96 in absolute value.

We repeat this for 12 different values of the reduced form error correlation. In Table 3 we report the coverage rate and the median and 0.10 quantile of the width of the estimated 95% confidence intervals.

Table 3: Coverage Rates of Conv. TSLS CI by Degree of Endogeneity

ρ	0.0	0.4	0.6	0.8	0.9	0.95	0.99
implied OLS	0.00	0.08	0.13	0.17	0.19	0.20	0.21
Real QOB							
Cov rate	0.95	0.95	0.96	0.95	0.95	0.95	0.95
Med Width 95% CI	0.09	0.08	0.07	0.06	0.05	0.05	0.05
0.10 quant Width	0.08	0.07	0.06	0.05	0.04	0.04	0.04
Random QOB							
Cov rate	0.99	1.00	1.00	0.98	0.92	0.82	0.53
Med Width 95% CI	1.82	1.66	1.45	1.09	0.79	0.57	0.26
0.10 quant Width	0.55	0.51	0.42	0.33	0.24	0.17	0.08

In this example, unless the reduced form correlations are very high, e.g., at least 0.95, with irrelevant instruments the conventional confidence intervals are wide and have good coverage.

The amount of endogeneity that would be required for the conventional confidence intervals to be misleading is higher than one typically encounters in cross-section settings.

Put differently, although formally conventional confidence intervals are not valid uniformly over the parameter space (e.g., Dufour, 1997), the subsets of the parameter space where results are substantively misleading may be of limited interest.

This in contrast to the case with many weak instruments where especially TSLS can be misleading in empirically relevant settings.

3.A Single Weak Instrument

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i,$$

$$X_i = \pi_0 + \pi_1 \cdot Z_i + \eta_i,$$

with $(\varepsilon_i, \eta_i) \perp\!\!\!\perp Z_i$, and jointly normal with covariance matrix Σ. The reduced form for the first equation is

$$Y_i = \alpha_0 + \alpha_1 \cdot Z_i + \nu_i,$$

where the parameter of interest is $\beta_1 = \alpha_1/\pi_1$. Let

$$\Omega = \mathbb{E}\left[\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix}'\right], \quad \text{and} \quad \Sigma = \mathbb{E}\left[\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}'\right],$$

Standard IV estimator:

$$\widehat{\beta}_1^{\text{IV}} = \frac{\frac{1}{N}\sum_{i=1}^N \left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right)}{\frac{1}{N}\sum_{i=1}^N \left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right)},$$

Concentration parameter:

$$\lambda = \pi_1^2 \cdot \sum_{i=1}^N \left(Z_i - \overline{Z}\right)^2 / \sigma_\eta^2.$$

Normal approximations for numerator and denominator are accurate:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^N \left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right) - \text{Cov}(Y_i, Z_i)\right) \approx \mathcal{N}\left(0, V(Y_i \cdot Z_i)\right),$$

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^N \left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right) - \text{Cov}(X_i, Z_I)\right) \approx \mathcal{N}\left(0, V(X_i \cdot Z_i)\right).$$

If $\pi_1 \neq 0$, as the sample size gets large, then the ratio will eventually be well approximated by a normal distribution as well.

However, if $\text{Cov}(X_i, Z_i) \approx 0$, the ratio may be better approximated by a Cauchy distribution, as the ratio of two normals centered close to zero.

3.B Staiger-Stock Asymptotics and Uniformity

Staiger and Stock investigate the distribution of the standard IV estimator under an alternative asymptotic approximation.

The standard asymptotics (strong instrument asymptotics in the SS terminology) is based on fixed parameters and the sample size getting large.

In their alternative asymptotic sequence SS model π_1 as a function of the sample size, $\pi_{1N} = c/\sqrt{N}$, so that the concentration parameter converges to a constant:

$$\lambda \longrightarrow c^2 \cdot V(Z_i).$$

SS then compare coverage properties of various confidence intervals under this (weak instrument) asymptotic sequence.

The importance of the SS approach is in demonstrating for any sample size there are values of the nuisance parameters such that the actual coverage is substantially away from the nominal coverage.

More recently the issue has therefore been reformulated as requiring confidence intervals to have asymptotically the correct coverage probabilities uniformly in the parameter space. See for a discussion from this perspective Mikusheva (2008).

Note that there **cannot** exist estimators that are consistent for β^* uniformly in the parameter space, since if $\pi_1 = 0$, β_1 is not identified. However, for testing there are generally confidence intervals that are uniformly valid, but they are not of the conventional form, that is, a point estimate plus or minus a constant times a standard error.

3.C Anderson-Rubin Confidence Intervals

Let the instrument $\tilde{Z}_i = Z_i - \overline{Z}$ be measured in deviations from its mean. Then define the statistic

$$S(\beta_1) = \frac{1}{N} \sum_{i=1}^{N} \tilde{Z}_i \cdot (Y_i - \beta_1 \cdot X_i).$$

Then, under the null hypothesis that $\beta_1 = \beta_1^*$, and conditional on the instruments, the statistic $\sqrt{N} \cdot S(\beta_1^*)$ has an exact normal distribution

$$\sqrt{N} \cdot S(\beta_1^*) \sim \mathcal{N}\left(0, \sum_{i=1}^{N} \tilde{Z}_i^2 \cdot \sigma_\varepsilon^2\right).$$

Anderson and Rubin (1949) propose basing tests for the null hypothesis

$$H_0: \quad \beta_1 = \beta_1^0, \qquad \text{against the alternative hypothesis} \quad H_a: \quad \beta_1 \neq \beta_1^0$$

on this idea, through the statistic

$$\mathsf{AR}\left(\beta_1^0\right) = \frac{N \cdot S(\beta_1^0)^2}{\sum_{i=1}^{N} \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix} \right)^{-1}.$$

A confidence interval can be based on this test statistic by inverting it:

$$\mathsf{CI}_{0.95}^{\beta_1} = \{\beta_1 \,|\mathsf{AR}(\beta_1) \leq 3.84\}$$

This interval can be equal to the whole real line.

3.D Anderson-Rubin with K instruments

The reduced form is

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i,$$

$S\left(\beta_1^0\right)$ is now normally distributed vector.
AR statistic with associated confidence interval:

$$\mathrm{AR}\left(\beta_1^0\right) = N \cdot S\left(\beta_1^0\right)' \left(\sum_{i=1}^{N} \tilde{Z}_i \cdot \tilde{Z}_i'\right)^{-1} S\left(\beta_1^0\right) \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)$$

$$\mathrm{CI}_{0.95}^{\beta_1} = \left\{\beta_1 \middle| \mathrm{AR}(\beta_1) \leq \mathcal{X}_{0.95}^2(K)\right\},$$

The problem is that this confidence interval can be empty because it simultaneously tests validity of instruments.

3.E Kleibergen Test

Kleibergen modfies AR statistic through

$$\tilde{S}\left(\beta_1^0\right) = \frac{1}{N} \sum_{i=1}^{N} \left(\tilde{Z}_i' \hat{\pi}_1(\beta_1^0)\right) \cdot \left(Y_i - \beta_1^0 \cdot X_i\right),$$

where $\hat{\pi}$ is the maximum likelihood estimator for π_1 under the restriction $\beta_1 = \beta_1^0$. The test is then based on the statistic

$$K\left(\beta_1^0\right) = \frac{N \cdot \tilde{S}(\beta_1^0)^2}{\sum_{i=1}^{N} \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

This has an approximate chi-squared distribution, and can be used to construct a confidence interval. Alternative: Moreira conditional likelihood ratio test.

4.A Many (Weak) Instruments

In this section we discuss the case with many weak instruments. The problem is both the bias in the standard estimators, and the misleadingly small standard errors based on conventional procedures, leading to poor coverage rates for standard confidence intervals in many situations.

Resampling methods such as bootstrapping do not solve these problems.

The literature has taken a number of approaches. Part of the literature has focused on alternative confidence intervals analogues to the single instrument case. In addition a variety of new point estimators have been proposed.

Recommendation: Use LIML, but adjust standard errors using Bekker correction.

4.B Bekker Asymptotics

Bekker (1995) derives large sample approximations for TSLS and LIML based on sequences where the number of instruments increases proportionally to the sample size.

He shows that TSLS is not consistent in that case.

LIML is consistent, but the conventional LIML standard errors are not valid. Bekker then provides LIML standard errors that are valid under this asymptotic sequence. Even with relatively small numbers of instruments the differences between the Bekker and conventional asymptotics can be substantial.

Bekker correction, single endogenous regressor:

$$Y_i = \beta_1' X_{1i} + \beta_2' X_{2i} + \varepsilon_i = \beta' X_i + \varepsilon_i,$$

$$X_{1i} = \pi_1' Z_{1i} + \pi_2' X_{2i} + \eta_i = \pi' Z_i + \eta_i.$$

Define the matrices $\mathbf{P}_Z$ and $\mathbf{M}_Z$ as:

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}', \quad \mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

Let σ^2 be the variance of ε_i, with consistent estimator $\hat{\sigma}^2$. The standard TSLS variance is

$$V_{tsls} = \hat{\sigma}^2 \cdot (\mathbf{X}\mathbf{P}_Z\mathbf{X})^{-1}.$$

Under the standard, fixed number of instrument asymptotics, the asymptotic variance for LIML is identical to that for TSLS, and so in principle we can use the same estimator. In practice researchers typically estimate the variance for LIML as

$$V_{liml} = \hat{\sigma}^2 \cdot \left(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda} \cdot \mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}.$$

To get Bekker's correction, we need a little more notation. Define

$$\Omega = \begin{pmatrix} \mathbf{Y} & \mathbf{X} \end{pmatrix} \mathbf{P}_Z \begin{pmatrix} \mathbf{Y} & \mathbf{X} \end{pmatrix} / N = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{pmatrix},$$

$$\Omega_{11} = \mathbf{Y}\mathbf{P}_Z\mathbf{Y}/N, \quad \Omega_{12} = \mathbf{Y}\mathbf{P}_Z\mathbf{X}/N, \quad \text{and} \quad \Omega_{22} = \mathbf{X}\mathbf{P}_Z\mathbf{X}/N.$$

$$\mathbf{A} = N \cdot \frac{\Omega_{12}'\Omega_{12} - \Omega_{22}\beta\Omega_{12} - \Omega_{12}'\beta'\Omega_{22} + \Omega_{22}\beta\beta'\Omega_{22}}{\Omega_{11} - 2\Omega_{12}\beta + \beta'\Omega_{22}\beta}.$$

Then:

$$V_{bekker} = \hat{\sigma}^2 \cdot \left(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda} \cdot \mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}$$

$$\times (\mathbf{X}\mathbf{P}_Z\mathbf{X} - \lambda \cdot \mathbf{A}) \cdot \left(\mathbf{X}\mathbf{P}_Z\mathbf{X} - \hat{\lambda} \cdot \mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}.$$

Recommended in practice

4.C Random Effects Estimators

Chamberlain and Imbens propose a random effects quasi maximum likelihood (REQML) estimator. They propose modelling the first stage coefficients π_k, for $k = 1, \ldots, K$, in the regression

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i = \pi_0 + \sum_{k=1}^{K} \pi_k \cdot Z_{ik} + \eta_i,$$

(after normalizing the instruments to have mean zero and unit variance,) as independent draws from a normal $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ distribution.

Assuming also joint normality for (ε_i, η_i), one can derive the likelihood function

$$\mathcal{L}(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2, \Omega).$$

In contrast to the likelihood function in terms of the original parameters $(\beta_0, \beta_1, \pi_0, \pi_1, \Omega)$, this likelihood function depends on a small set of parameters, and a quadratic approximation to its logarithms is more likely to be accurate.

CI discuss some connections between the REQML estimator and LIML and TSLS in the context of this parametric set up. First they show that in large samples, with a large number of instruments, the TSLS estimator corresponds to the restricted maximum likelihood estimator where the variance of the first stage coefficients is fixed at a large number, or $\sigma_\pi^2 = \infty$:

$$\hat{\beta}_{\text{TSLS}} \approx \arg \max_{\beta_0, \beta_1, \pi_0, \mu_\pi} = L(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2 = \infty, \Omega).$$

From a Bayesian perspective, TSLS corresponds approximately to the posterior mode given a flat prior on all the parameters, and thus puts a large amount of prior mass on values of the parameter space where the instruments are jointly powerful.

In the special case where we fix $\mu_\pi = 0$, and Ω is known, and the random effects specification applies to all instruments, CI show that the REQML estimator is identical to LIML.

However, like the Bekker asymptotics, the REQML calculations suggests that the standard LIML variance is too small: the variance of the REQML estimator is approximately equal to the standard LIML variance times

$$1 + \sigma_\pi^{-2} \cdot \left(\begin{pmatrix} 1 \\ \beta_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \right)^{-1}.$$

This is similar to the Bekker adjustment.

4.E Flores' Simulations

In one of the more extensive simulation studies Flores-Lagunes (2007) reports results comparing TSLS, LIML, Fuller, Bias corrected versions of TSLS, LIML and Fuller, a Jacknife version of TSLS (Hahn, Hausman and Kuersteiner), and the REQML estimator, in settings with 100 and 500 observations, and 5 and 30 instruments for the single endogenous variable. (Does not include LIML with Bekker standard errors.)

He concludes that "our evidence indicates that the random-effects quasi-maximum likelihood estimator outperforms alternative estimators in terms of median point estimates and coverage rates."

Part II: Generalized Method of Moments and Empirical Likelihood

1. Introduction
2. Generalized Method of Moments Estimation
3. Empirical Likelihood
4. Computational Issues
5. A Dynamic Panel Data Model

1. Introduction

GMM has provided a very influential framework for estimation since Hansen (1982). Many models and estimators fit in.

In the case with over-identification the traditional approach is to use a two-step method with estimated weight matrix. Can perform poorly in settings with high degree of overidentification.

In such settings Empirical Likelihood provides attractive alternative with higher order bias properties, and liml-like advantages.

The choice between various EL-type estimators is less important. Computationally the estimators are only marginally more demanding. Most effective seems to be to concentrate out Lagrange multipliers.

2. Generalized Method of Moments Estimation

Generic form of the GMM estimation problem: The parameter vector θ^* is a K dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$.

The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector valued function such that

$$\mathbb{E}\left[\psi(Z, \theta^*)\right] = 0, \quad \text{and} \quad \mathbb{E}\left[\psi(Z, \theta)\right] \neq 0, \text{ for all } \theta \neq \theta^*$$

The researcher has available an independent and identically distributed random sample $Z_1, Z_2, \ldots, Z_N$. We are interested in the properties of estimators for θ^* in large samples.

Example: A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T.

This is a stylized version of the type of panel data models studied in Keane and Runkle (1992), Chamberlain (1992), and Blundell and Bond (1998). This specific model has previously been studied by Bond, Bowsher, and Windmeijer (2001).

One can construct moment functions by differencing and using lags as instruments, as in Arellano and Bond (1991), and Ahn and Schmidt, (1995):

$$\psi_{1t}(Y_{i1},\ldots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot ((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2})).$$

This leads to $t-2$ moment functions for each value of $t = 3,\ldots,T$, leading to a total of $(T-1)\cdot(T-2)/2$ moments, with only a single parameter (θ).

In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1},\ldots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

GMM: Estimation

In the just-identified case (dim(ψ) =dim(θ)), one can estimate θ^* by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \widehat{\theta}_{\text{gmm}}). \tag{1}$$

Under regularity conditions solutions will be unique in large samples and consistent for θ^*. If $M > K$ there is in general there will be no solution to (1).

Hansen's solution was to minimize the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i,\theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N} \psi(z_i,\theta)\right],$$

for some positive definite $M \times M$ symmetric matrix C (which if $M = K$ still leads to a $\widehat{\theta}$ that solves the equation (1).

GMM: Large Sample Properties

Under regularity conditions the minimand $\widehat{\theta}_{\text{gmm}}$ has the following large sample properties:

$$\widehat{\theta}_{\text{gmm}} \xrightarrow{p} \theta^*,$$

$$\sqrt{N}(\widehat{\theta}_{\text{gmm}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1}\Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}),$$

where

$$\Delta = \mathbb{E}\left[\psi(Z_i,\theta^*)\psi(Z_i,\theta^*)'\right] \quad \text{and} \quad \Gamma = \mathbb{E}\left[\frac{\partial}{\partial\theta'}\psi(Z_i,\theta^*)\right].$$

In the just–identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial the asymptotic covariance matrix reduces to $(\Gamma'\Delta^{-1}\Gamma)^{-1}$.

GMM: Optimal Weight Matrix

In the overidentified case with $M > K$ the choice of the weight matrix C is important.

The optimal choice for C in terms of minimizing the asymptotic variance is in this case the inverse of the covariance of the moments, Δ^{-1}.

Then:

$$\sqrt{N}(\widehat{\theta}_{\text{gmm}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma'\Delta^{-1}\Gamma)^{-1}). \tag{2}$$

This estimator is not feasible because Δ^{-1} is unknown.

The feasible solution is to obtain an initial consistent, but generally inefficient, estimate of θ^* and then can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})'\right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\hat{\Delta}^{-1},N}(\theta)$.

The resulting estimator $\hat{\theta}_{\text{gmm}}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Compare to TSLS having the same asymptotic distribution as estimator with optimal instrument.

3. Empirical Likelihood

Consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. The natural choice for estimating the dsitribution function is the empirical distribution, that puts weight $1/N$ on each of the N sample points.

Suppose we also know that $\mathbb{E}[Z] = 0$. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\hat{F}_{emp}}[Z] = \sum z_i/N \neq 0$.

The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\hat{F}$ does satisfy the restriction.

The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i, \quad \text{for } 0 \leq \pi_i \leq 1, \ \sum_{i=1}^{N} \pi_i = 1$$

The empirical likelihood estimator for the distribution function is, given $\mathbb{E}[Z] = 0$,

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i \cdot z_i = 0.$$

Without the second restriction the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied.

This leads to

$$\hat{\pi}_i = 1/(1 + t \cdot z_i) \quad \text{where } t \text{ solves } \sum_{i=1}^{N} \frac{z_i}{1 + t \cdot z_i} = 0,$$

EL: The General Case

More generally, in the over-identified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta,\pi} \sum_{i=1}^{N} \ln \pi, \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0.$$

This is equivalent, to first order asymptotics, to the two-step GMM estimator.

For many purposes the empirical likelihood has the same properties as a parametric likelihood function. (Qin and Lawless, 1994; Imbens, 1997; Kitamura and Stutzer, 1997).

EL: Cressie-Read Discrepancy Statistics

Define

$$I_\lambda(p,q) = \frac{1}{\lambda \cdot (1+\lambda)} \sum_{i=1}^{N} p_i \left[\left(\frac{p_i}{q_i} \right)^\lambda - 1 \right].$$

and solve

$$\min_{\pi,\theta} I_\lambda(\iota/N, \pi) \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1, \quad \text{and} \quad \sum_{i=1}^{N} \pi_i \cdot \psi(z_i,\theta) = 0.$$

The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Empirical Likelihood is special case with $\lambda = \longrightarrow 0$.

EL: Generalized Empirical Likelihood

Smith (1997), Newey and Smith (1994) considers a more general class of estimators. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$, $g'(0) = 1$, consider the saddle point problem

$$\max_\theta \min_t \sum_{i=1}^{N} g(t'\psi(z_i,\theta)).$$

This representation is attractive from a computational perspective, as it reduces the dimension of the optimization problem to $M + K$ rather than a constrained optimization problem of dimension $K + N$ with $M + 1$ restrictions.

There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie-Read representation.

EL: Special cases, Continuously Updating Estimator

$\lambda = -2$.

This case was originally proposed by Hansen, Heaton and Yaron (1996) as the solution to

$$\min_\theta \frac{1}{N}[\sum_{i=1}^{N} \psi(z_i,\theta)]' \cdot \left[\frac{1}{N} \sum_{i=1}^{N} \psi(z_i,\theta)\psi(z_i,\theta)' \right]^{-1} \cdot [\sum_{i=1}^{N} \psi(z_i,\theta)],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments.

Newey and Smith (2004) pointed out that this estimator fits in the Cressie-Read class.

EL: Special cases, Exponential Tilting Estimator

$\lambda \longrightarrow -1$.

The second case is the exponential tilting estimator with $\lambda \to -1$ (Imbens, Spady and Johnson, 1998), whose objective function is equal to the empirical likelihood objective funtion with the role of π and ι/N reversed.

It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^{N} \pi_i \cdot \ln \pi_i \quad \text{subject to} \quad \sum_{i=1}^{N} \pi_i = 1, \quad \text{and} \quad \sum_{i=1}^{N} \pi_i \cdot \psi(z_i,\theta) = 0.$$

Comparison of GEL Estimators

Little known in general.

EL ($\lambda = 0$) has higher order bias properties (NS), but implicit probabilities can get large.

CUE ($\lambda = -2$) tends to have more outliers

ET ($\lambda = -1$) computationally stable.

Comparison GMM and EL: Newey and Smith, 2004

Comparison of expectation of second order term:

$$\widehat{\theta} = \theta_0 + \frac{A}{\sqrt{N}} + \frac{B}{N} + o_p\left(N^{-1}\right)$$

Bias is $\mathbb{E}[B]$.

$$\text{Bias}\left(\widehat{\theta}_{GEL}\right) = B_I + \left(1 + \frac{\rho}{2}\right) B_\Omega \qquad \rho_{EL} = -2$$

$$\text{Bias}\left(\widehat{\theta}_{GMM}\right) = B_I + B_\Omega + B_G + B_W$$

B_Ω comes from third moment of moments: $\mathbb{E}[\psi^3]$

B_W comes from estimating weight matrix.

B_G comes from correlation between $\partial\psi/\partial\theta$ and ψ.

4. Computational Issues: Concentrating out the Lagrange Multipliers

Mittelhammer, Judge and Schoenberg (2001) suggest concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. This appears to work well.

Concentrating out the probabilities π_i can be done analytically.

Although it is not in general possible to solve for the Lagrange multipliers t analytically for given θ it is easy to numerically solve for t. E.g., in the exponential tilting case, solve

$$\min_t \sum_{i=1}^N \exp(t'\psi(z_i, \theta)).$$

This function is strictly convex as a function of t, with easy-to-calculate first and second derivatives.

After solving for $t(\theta)$, one can solve

$$\max_\theta \sum_{i=1}^N \exp(t(\theta)'\psi(z_i, \theta)).$$

Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ.

The function $t(\theta)$ has analytic derivatives with respect to θ.

5. Dynamic Panel Data Model (as described before)

To get a sense of the finite sample properties of the empirical likelihood estimators we compare two-step GMM and one of the EL estimators (exponential tilting) in the context of a panel data model

The model is

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$.

Moments:

$$\psi_{1t}(Y_{i1}, \ldots, Y_{iT}, \theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot ((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2})).$$

This leads to $(T-1) \cdot (T-2)/2$ moments.

Additional $T-2$ moments:

$$\psi_{2t}(Y_{i1}, \ldots, Y_{iT}, \theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

Note that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. So, potentially substantial difference between estimators.

We report some simulations for a data generating process with parameter values estimated on data from Abowd and Card (1989) taken from the PSID. See also Card (1994).

This data set contains earnings data for 1434 individuals for 11 years. The individuals are selected on having positive earnings in each of the eleven years, and we model their earnings in logarithms. We focus on estimation of the autoregressive coefficient θ.

Using the Abowd-Card data we estimate θ and the variance of the fixed effect and the idiosyncratic error term. The latter two are estimated to be around 0.3. We use $\theta = 0.5$ and $\theta = 0.9$ in the simulations. The first is comparable to the value estimated from the Abowd-Card data.

$\theta = 0.5$	Number of time periods				
	3	4	6	7	9
Two-Step GMM					
median bias	-0.00	0.00	-0.00	-0.00	0.00
relative median bias	-0.07	0.01	-0.06	-0.08	0.09
median absolute error	0.05	0.03	0.01	0.01	0.01
coverage rate 90% ci	0.91	0.88	0.91	0.91	0.89
covarage rate 95% ci	0.95	0.94	0.95	0.96	0.95
Exponential Tilting					
median bias	-0.00	-0.00	-0.00	-0.00	0.00
relative median bias	-0.04	-0.02	-0.09	-0.07	0.02
median absolute error	0.05	0.03	0.01	0.01	0.01
coverage rate 90% ci	0.90	0.87	0.90	0.92	0.90
covarage rate 95% ci	0.95	0.94	0.96	0.95	0.95

$\theta = 0.9$	Number of time periods				
	3	4	6	7	9
Two-Step GMM					
median bias	-0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.08	0.03	0.08
median absolute error	0.04	0.03	0.02	0.02	0.01
coverage rate 90% ci	0.88	0.85	0.80	0.80	0.78
covarage rate 95% ci	0.92	0.91	0.87	0.85	0.86
Exponential Tilting					
median bias	0.00	0.00	-0.00	0.00	-0.00
relative median bias	0.04	0.09	-0.00	0.01	-0.02
median absolute error	0.05	0.03	0.02	0.02	0.01
coverage rate 90% ci	0.87	0.86	0.86	0.88	0.87
covarage rate 95% ci	0.91	0.90	0.91	0.93	0.91

AEA Lectures **Chicago, IL, January 2012**

Lecture 9, Tuesday, Jan 10th, am

Partial Identification

1. Introduction

Traditionally in constructing statistical or econometric models researchers look for models that are *(point-)identified*: given a large (infinite) data set, one can infer without uncertainty what the values are of the objects of interest, the estimands. Even though the fact that a model is identified does not necessarily imply that we do well in finite samples, it would appear that a model where we cannot learn the parameter values even in infinitely large samples would not be very useful. Traditionally therefore researchers have stayed away from models that are not (point-)identified, often adding assumptions beyond those that could be justified using substantive arguments. However, it turns out that even in cases where we cannot learn the value of the estimand *exactly* in large samples, in many cases we can still learn a fair amount, even in finite samples. A research agenda initiated by Manski (an early paper is Manski (1990), monographs include Manski (1995, 2003)), referred to as *partial identification*, or earlier as *bounds*, and more recently adopted by a large number of others, notably Tamer in a series papers (Haile and Tamer, 2003, Ciliberto and Tamer, 2007; Aradillas-Lopez and Tamer, 2007), has taken this perspective. In this lecture we focus primarily on a number of examples to show the richness of this approach. In addition we discuss some of the theoretical issues connected with this literature, and some practical issues in implementation of these methods.

The basic set up we adopt is one where we have a random sample of units from some population. For the typical unit, unit i, we observe the value of a vector of variables Z_i. Sometimes it is useful to think of there being in the background a latent variable variable W_i. We are interested in some functional θ of the joint distribution of Z_i and W_i, but, not observing W_i for any units, we may not be able to learn the value of θ even in infinite samples because the estimand cannot be written as a functional of the distribution of Z_i alone. The

three key questions are (i) what we can learn about θ in large samples (identification), (ii) how do we estimate this (estimation), and (iii) how do we quantify the uncertainty regarding θ (inference).

The solution to the first question will typically be a set, the *identified set*. Even if we can characterize estimators for these sets, computing them can present serious challenges. Finally, inference involves challenges concerning uniformity of the coverage rates, as well as the question whether we are interested in coverage of the entire identified set or only of the parameter of interest.

There are a number of cases of general interest. I will discuss two leading cases in more detail. In the first case the focus is on a scalar, with the identified set equal to an interval with lower and upper bound a smooth, $\sqrt{N}$-estimable functional of the data. A second case of interest is that where the information about the parameters can be characterized by moment restrictions, often arising from revealed preference comparisons between utilities at actions taken and actions not taken. I refer to this as the generalized inequality restrictions (GIR) setting. This set up is closely related to the generalized method of moments framework.

2. Partial Identification: Examples

Here we discuss a number of examples to demonstrate the richness of the partial identification approach.

2.1 Missing Data

This is a basic example, see e.g., Manski (1990), and Imbens and Manski (2004). It is substantively not very interesting, but it illustrates a lot of the basic issues. Suppose the observed variable is the pair $Z_i = (D_i, D_i \cdot Y_i)$, and the unobserved variable is $W_i = Y_i$. D_i is a binary variable. This corresponds to a missing data case. If $D_i = 1$, we observe Y_i, and if $D_i = 0$ we do not observe Y_i. We always observe the missing data indicator D_i. We assume the quantity of interest is the population mean $\theta = \mathbb{E}[Y_i]$.

In large samples we can learn $p = \mathbb{E}[D_i]$ and $\mu_1 = \mathbb{E}[Y_i|D_i = 1]$. The data contain no

information about $\mu_0 = \mathbb{E}[Y_i|D_i = 0]$. It can be useful, though not always possible, to write the estimand in terms of parameters that are point-identified and parameters that the data are not informative about. In this case we can do so:

$$\theta = p \cdot \mu_1 + (1-p) \cdot \mu_0.$$

Since even in large samples we learn nothing about μ_0, it follows that without additional information there is no limit on the range of possible values for θ. Even if p is very close to 1, this small probability that $D_i = 0$ combined with the possibility that μ_0 is very large or very small allows for a wide range of values for θ.

Now suppose we know that the variable of interest is binary: $Y_i \in \{0, 1\}$. Then natural (not data-informed) lower and upper bounds for μ_0 are 0 and 1 respectively. This implies bounds on θ:

$$\theta \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = [p \cdot \mu_1, p \cdot \mu_1 + (1-p)] \,.$$

These bounds are *sharp*, in the sense that without additional information we can not improve on them. Formally, for all values θ in $[\theta_{\text{LB}}, \theta_{\text{UB}}]$, we can find a joint distribution of (Y_i, W_i) that is consistent with the joint distribution of the observed data and with θ. Even if Y is not binary, but has some natural bounds, we can obtain potentially informative bounds on θ.

We can also obtain informative bounds if we modify the object of interest a little bit. Suppose we are interested in quantiles of the distribution of Y_i. To make this specific, suppose we are interested in the median of Y_i, $\theta_{0.5} = \text{med}(Y_i)$. The largest possible value for the median arises if all the missing value of Y_i are large. Define $q_\tau(Y_i|D_i = d)$ to be the τ quantile of the conditional distribution of Y_i given $D_i = d$. Then the median cannot be larger than $q_{1/(2p)}(Y_i|D_i = 1)$ because even if all the missing values were large, we know that at least $p \cdot (1/(2p)) = 1/2$ of the units have a value less than or equal to $q_{1/(2p)}(Y_i|D_i = 1)$. Similarly, the smallest possible value for the median corrresponds to the case where all the

missing values are small, leading to a lower bound of $q_{(2p-1)/(2p)}(Y_i|D_i=1)$. Then, if $p>1/2$, we can infer that the median must satisfy

$$\theta_{0.5}\in[\theta_{\rm LB},\theta_{\rm UB}]=\left[q_{(2p-1)/(2p)}(Y_i|D_i=1),q_{1/(2p)}(Y_i|D_i=1)\right],$$

and we end up with a well defined, and, depending on the data, more or less informative identified interval for the median. If fewer than 50% of the values are observed, or $p<1/2$, then we cannot learn anything about the median of Y_i without additional information (for example, a bound on the values of Y_i), and the interval is $(-\infty,\infty)$. More generally, we can obtain bounds on the τ quantile of the distribution of Y_i, equal to

$$\theta_\tau\in[\theta_{\rm LB},\theta_{\rm UB}]=\left[q_{(\tau-(1-p))/p}(Y_i|D_i=1),q_{\tau/p}(Y_i|D_i=1)\right].$$

which is bounded if the probability of Y_i being missing is less than $\min(\tau,1-\tau)$.

2.2 Returns to Schooling

Manski and Pepper (2000, MP) are interested in estimating returns to schooling. They start with an individual level response function $Y_i(w)$, where $w\in\{0,1,\ldots,20\}$ is years of schooling. Let

$$\Delta(s,t)=\mathbb{E}[Y_i(t)-Y_i(s)],$$

be the difference in average outcomes (log earnings) given t rather than s years of schooling. Values of $\Delta(s,t)$ at different combinations of (s,t) are the object of interest. Let W_i be the actual years of school, and $Y_i=Y_i(W_i)$ be the actual log earnings. If one makes an unconfoundedness type assumption that

$$Y_i(w)\ \perp\!\!\!\perp\ W_i\ \Big|\ X_i,$$

for some set of covariates, one can estimate $\Delta(s,t)$ consistently given some support conditions. MP relax this assumption. Dropping this assumption entirely without additional

assumptions one can derive the bounds using the missing data results in the previous section. In this case most of the data would be missing, and the bounds would be wide. More interestingly MP focus on a number of alternative, weaker assumptions, that do not allow for point-identification of $\Delta(s,t)$, but that nevertheless may be able to narrow the range of values consistent with the data to an informative set. One of their assumptions requires that increasing education does not lower earnings:

Assumption 1 (Monotone Treatment Response)
If $w' \geq w$, *then* $Y_i(w') \geq Y_i(w)$.

Another assumption states that, on average, individuals who choose higher levels of education would have higher earnings at each level of education than individuals who choose lower levels of education.

Assumption 2 (Monotone Treatment Selection)
If $w'' \geq w'$, *then for all* w, $\mathbb{E}[Y_i(w)|W_i = w''] \geq \mathbb{E}[Y_i(w)|W_i = w']$.

Both assumptions are consistent with many models of human capital accumulation. They also address the main concern with the exogenous schooling assumption, namely that higher ability individuals who would have had higher earnings in the absence of more schooling, are more likely to acquire more schooling.

Under these two assumptions, the bound on the average outcome given w years of schooling is

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \geq w) + \sum_{v<w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v)$$

$$\leq \mathbb{E}[Y_i(w)] \leq$$

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \leq w) + \sum_{v>w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v).$$

Using data from the National Longitudinal Study of Youth MP a point estimator for the upper bound on the the returns to four years of college, $\Delta(12,16)$ to be 0.397, with a 0.95 upper quantile of 0.450. Translated into an average yearl returns this gives us 0.099, which is in fact lower than some estimates that have been reported in the literature. This analysis suggests that the upper bound is in this case reasonably informative, given a remarkably weaker set of assumptions.

2.3 Changes in Inequality and Selection

There is a large literature on the changes in the wage distribution and the role of changes in the returns to skills that drive these changes. One concern is that if one compares the wage distribution at two points in time, any differences may be partly or wholly due to differences in the composition of the workforce. Blundell, Gosling, Ichimura, and Meghir (2007, BGHM) investigate this using bounds. They study changes in the wage distribution in the United Kingdom for both men and women. Even for men at prime employment ages employment in the late nineties is less than 0.90, down from 0.95 in the late seventies. The concern is that the 10% who do not work are potentially different, both from those who work, as well as from those who did not work in the seventies, corrupting comparisons between the wage distributions in both years. Traditionally such concerns may have been ignored by implicitly assuming that the wages for those not working are similar to those who are working, possibly conditional on some observed covariates, or they may have been addressed by using selection models. The type of selection models used ranges from very parametric models of the type originally developed by Heckman (1978), to semi- and non-parametric versions of this (Heckman, 1990). The concern that BGHM raise is that those selection models rely on assumptions that are difficult to motivate by economic theory. They investigate what can be learned about the changes in the wage distributions without the final, most controversial assumptions of those selection models.

BGHM focus on the interquartile range as their measure of dispersion in the wage distribution. As discussed in Section 2.1, this is convenient, because bounds on quantiles often exist in the presence of missing data. Let $F_{Y|X}(y|x)$ be the distribution of wages condi-

tional on some characteristics X. This is assumed to be well defined irrespective of whether an individual works or not. However, if an individual does not work, Y_i is not observed. Let D_i be an indicator for employment. Then we can estimate the conditional wage distribution given employment, $F_{Y|X,D}(y|x,d=1)$, as well as the probability of employment, $p(x) = \text{pr}(D_i = 1|X_i = x)$. This gives us tight bounds on the (unconditional on employment) wage distribution

$$F_{Y|X,D}(y|x,d=1)\cdot p(x) \leq F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=1)\cdot p(x) + (1-p(x)).$$

We can convert this to bounds on the τ quantile of the conditional distribution of Y_i given $X_i = x$, denoted by $q_\tau(x)$:

$$q_{(\tau-(1-p(x)))/p(x)}(Y_i|D_i=1) \leq q_\tau(x) \leq q_{\tau/p(x)}(Y_i|D_i=1),$$

Then this can be used to derive bounds on the interquartile range $q_{0.75}(x) - q_{0.25}(x)$:

$$q_{(0.75-(1-p(x)))/p(x)}(Y_i|D_i=1) - q_{0.25/p(x)}(Y_i|D_i=1)$$

$$\leq q_{0.75}(x) - q_{0.25}(x) \leq$$

$$q_{(0.25-(1-p(x)))/p(x)}(Y_i|D_i=1) - q_{0.75/p(x)}(Y_i|D_i=1).$$

So far this is just an application of the missing data bounds derived in the previous section. What makes this more interesting is the use of additional information short of imposing a full selection model that would point identify the interquartile range. The first assumption BGHM add is that of stochastic dominance of the wage distribution for employed individuals:

$$F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=0).$$

One can argue with this stochastic dominance assumption, but within groups homogenous in background characteristics including education, it may be reasonable. This assumption tightens the bounds on the distribution function to:

$$F_{Y|X,D}(y|x,d=1) \leq F_{Y|X,D}(y|x,d=1) \leq$$

$$F_{Y|X,D}(y|x,d=1)\cdot p(x) + (1-p(x)).$$

Another assumption BGHM consider is a modification of an instrumental variables assumption that an observed covariate Z is excluded from the wage distribution:

$$F_{Y|X,Z}(y|X=x,Z=z) = F_{Y|X,Z}(y|X=x,Z=z'), \qquad \text{for all } x,z,z'.$$

This changes the bounds on the distribution function to:

$$\max_z F_{Y|X,Z,D}(y|x,z,d=1)\cdot p(x,z)$$

$$\leq F_{Y|X,D}(y|x) \leq$$

$$\min_z F_{Y|X,Z,D}(y|x,z,d=1)\cdot p(x) + (1-p(x)).$$

(An alternative weakening of the standard instrumental variables assumption is in Hotz, Mullin and Sanders (1997), where a valid instrument exists, but is not observed directly.)

Such an instrument may be difficult to find, and BGHM argue that it may be easier to find a covariate that affects the wage distribution in one direction, using a monotone instrumental variables restriction suggested by Manski and Pepper (2000):

$$F_{Y|X,Z}(y|X=x,Z=z) \leq F_{Y|X,Z}(y|X=x,Z=z'), \qquad \text{for all } x,z<z'.$$

This discussion is somewhat typical of what is done in empirical work in this area. A number of assumptions are considered, with the implications for the bounds investigated. The results lay out part of the mapping between the assumptions and the bounds.

2.4 Random Effects Panel Data Models with Initial Condition Problems

Honoré and Tamer (2006) study dynamic random effects panel data models. We observe $(X_{i1}, Y_{i1}, \ldots, X_{iT}, Y_{iT})$, for $i = 1, \ldots, N$. The time dimension T is small relative to the cross-section dimension N. Large sample approximations are based on fixed T and large N. Inference would be standard if we specified a parametric model for the (components of the) conditional distribution of $(Y_{i1}, \ldots, Y_{iT})$ given $(X_{i1}, \ldots, X_{iT})$. In that case we could use maximum likelihood methods. However, it is difficult to specify this conditional distribution directly. Often we start with a model for the evolution of Y_{it} in terms of the present and past covariates and its lags. As an example, consider the model

$$Y_{it} = 1\{X_{it}'\beta + Y_{it-1}\gamma + \alpha_i + \epsilon_{it} \geq 0\},$$

with the ϵ_{it} independent over time and individuals, and normally distributed, $\epsilon_{it} \sim \mathcal{N}(0,1)$. The object of interest is the parameter governing the dynamics, γ. This model gives us the conditional distribution of $Y_{i2}, \ldots, Y_{iT}$ given Y_{i1}, α_i and given $X_{i1}, \ldots, X_{iT}$. Suppose we also postulate a parametric model for the random effects α_i:

$$\alpha | X_{i1}, \ldots, X_{iT} \sim G(\alpha|\theta),$$

(so in this case α_i is independent of the covariates). Then the model is (almost) complete, in the sense that we can almost write down the conditional distribution of $(Y_{i1}, \ldots, Y_{iT})$ given $(X_{i1}, \ldots, X_{iT})$. All that is missing is the conditional distribution of the initial condition:

$$p(Y_{i1}|\alpha_i, X_{i1}, \ldots, X_{iT}).$$

This is a difficult distribution to specify. One could directly specify this distribution, but one might want it to be internally consistent across different number of time periods, and that makes it awkward to choose a functional form. See for general discussions of this initial conditions problem Wooldridge (2002). Honoré and Tamer investigate what can be learned about γ without making parametric assumptions about this distribution. From the literature

it is known that in many cases γ is not point-identified (for example, the case with $T \leq 3$, no covariates, and a logistic distribution for ϵ_{it}). Nevertheless, it may be that the range of values of γ consistent with the data is very small, and it might reveal the sign of γ.

Honoré and Tamer study the case with a discrete distribution for α, with a finite and known set of support points. They fix the support to be $-3, -2.8, \ldots, 2.8, 3$, with unknown probabilities. Given that the ϵ_{it} are standard normal, this is very flexible. In a computational exercise they assume that the true probabilities make this discrete distribution mimic the standard normal distribution. In addition they set $\Pr(Y_{i1} = 1|\alpha_i) = 1/2$. In the case with $T = 3$ they find that the range of values for γ consistent with the data generating process (the identified set) is very narrow. If γ is in fact equal to zero, the width of the set is zero. If the true value is $\gamma = 1$, then the width of the interval is approximately 0.1. (It is largest for γ close to, but not equal to, -1.) See Figure 1, taken from Honoré and Tamer (2006).

The Honoré-Tamer analysis, in the context of the literature on initial conditions problems, shows very nicely the power of the partial identification approach. A problem that had been viewed as essentially intractable, with many non-identification results, was shown to admit potentially precise inferences despite these non-identification results.

2.5 Auction Data

Haile and Tamer (2003, HT from hereon), in what is one of the most influential applications of the partial identification approach, study English or oral ascending bid auctions. In such auctions bidders offer increasingly higher prices until only one bidder remains. HT focus on a symmetric independent private values model. In auction t, for $t = 1, \ldots, T$, bidder i has a value ν_{it}, drawn independently from the value for bidder j. Large sample results refer to the number of auctions getting large. HT assume that the value distribution is the same in each auction (after adjusting for observable auction characteristics). A key object of interest, is the value distribution. Given that one can derive other interesting objects, such as the optimal reserve price.

One can imagine a set up where the researcher observes, as the price increases, for each

bidder whether that bidder is still participating in the auction. (Milgrom and Weber (1982) assume that each bidder continuously confirms their participation by holding down a button while prices rise continuously.) In that case one would be able to infer for each bidder their valuation, and thus directly estimate the value distribution.

This is not what is typically observed. Instead of prices rising continuously, there are jumps in the bids, and for each bidder we do not know at any point in time whether they are still participating unless they subsequently make a higher bid. HT study identification in this, more realistic, setting. They assume that no bidder ever bids more than their valuation, and that no bidder will walk away and let another bidder win the auction if the winning bid is lower than their own valuation. Under those two assumptions, HT show that one can derive bounds on the value distribution.

One set of bounds they propose is as follows. Let the highest bid for participant i in auction t be b_{it}. The number of participants in auction t is n_t. Ignoring any covariates, let the distribution of the value for individual i, ν_{it}, be $F_\nu(v)$. This distribution function is the same for all auctions. Let $F_b(b) = \Pr(b_{it} \leq b)$ be the distribution function of the bids (ignoring variation in the number of bidders by auction). This distribution can be estimated because the bids are observed. The winning bid in auction t is $B_t = \max_{i=1,\ldots,n_t} b_{it}$. First HT derive an upper bound on the distribution function $F_\nu(v)$. Because no bidder ever bids more than their value, it follows that $b_{it} \leq \nu_{it}$. Hence, without additional assumptions,

$$F_\nu(v) \leq F_b(v), \qquad \text{for all } v.$$

For a lower bound on the distribution function one can use the fact that the second highest of the values among the n participants in auction t must be less than or equal to the winning bid. This follows from the assumption that no participant will let someone else win with a bid below their valuation. Let $F_{\nu,m:n}(v)$ denote the mth order statistic in a random sample of size n from the value distribution, and let $F_{B,n:n}(b)$ denote the distribution of the

winning bid in auctions with n participants. Then

$$F_{B,n:n}(v) \leq F_{\nu,n-1:n}(v).$$

The distribution of the any order statistic is monotonically related to the distribution of the parent distribution, and so a lower bound on $F_{\nu,n-1:n}(v)$ implies a lower bound on $F_\nu(v)$.

HT derive tighter bounds based on the information in other bids and the inequalities arising from the order statistics, but the above discussion illustrates the point that outside of the Milgrom-Weber button auction model one can still derive bounds on the value distribution in an English auction even if one cannot point-identify the value distribution. If in fact the highest bid for each individual was equal to their value (other than for the winner for whom the bid is equal to the second highest value), the bounds would collaps and point-identification would be obtained.

2.6 Entry Models and Inequality Conditions

Recently a number of papers has studied entry models in settings with multiple equilibria. In such settings traditionally researchers have added *ad hoc* equilbrium selection mechanisms. In the recent literature a key feature is the avoidance of such assumptions, as these are often difficult to justify on theoretical grounds. Instead the focus is on what can be learned in the absence of such assumptions. In this section I will discuss some examples from this literature. An important feature of these models is that they often lead to inequality restrictions, where the parameters of interest θ satisfy

$$\mathbb{E}[\psi(Z,\theta)] \geq 0,$$

for known $\psi(z,\theta)$. This relates closely to the standard (Hansen, 1983) generalized method of moments (GMM) set up where the functions $\psi(Z,\theta)$ would have expectation equal to zero at the true values of the parameters. We refer to this as the generalized inequality restrictions (GIR) form. These papers include Pakes, Porter, Ho, and Ishii (2006), Cilberto and Tamer (2004, CM from hereon), Andrews, Berry and Jia (2004). Here I will discuss a simplified

version of the CM model. Suppose two firms, A and B, contest a set of markets. In market m, $m = 1, \ldots, M$, the profits for firms A and B are

$$\pi_{Am} = \alpha_A + \delta_A \cdot d_{Bm} + \varepsilon_{Am}, \qquad \text{and} \quad \pi_{Bm} = \alpha_B + \delta_B \cdot d_{Am} + \varepsilon_{Bm}.$$

where $d_{Fm} = 1$ if firm F is present in market m, for $F \in \{A, B\}$, and zero otherwise. The more realistic model CM consider also includes observed market and firm characteristics. Firms enter market m if their profits in that market are positive. Firms observe all components of profits, including those that are unobserved to the econometrician, $(\varepsilon_{Am}, \varepsilon_{Bm})$, and so their decisions satisfy:

$$d_{Am} = 1\{\pi_{Am} \geq 0\}, \qquad d_{Bm} = 1\{\pi_{Bm} \geq 0\}. \tag{1}$$

(Pakes, Porter, Ho, and Ishii allow for incomplete information where expected profits are at least as high for the action taken as for actions not taken, given some information set.) The unobserved (to the econometrician) components of profits, ε_{Fm}, are independent accross markets and firms. For ease of exposition we assume here that they have a normal $\mathcal{N}(0, 1)$ distribution. (Note that we only observe indicators of the sign of profits, so the scale of the unobserved components is not relevant for predictions.) The econometrician observes in each market only the pair of indicators d_A and d_B. We focus on the case where the effect of entry of the other firm on a firm's profits, captured by the parameters δ_A and δ_B is negative, which is the case of most economic interest.

An important feature of this model is that given the parameters $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, for a given set of $(\varepsilon_{Am}, \varepsilon_{Bm})$ there is not necessarily a unique solution (d_{Am}, d_{Bm}). For pairs of values $(\varepsilon_{Am}, \varepsilon_{Bm})$ such that

$$-\alpha_A < \varepsilon_A \leq -\alpha_A - \delta_A, \qquad -\alpha_B < \varepsilon_B \leq -\alpha_B - \delta_B,$$

both $(d_A, d_B) = (0, 1)$ and $(d_A, d_B) = (1, 0)$ satisfy the profit maximization condition (1). In the terminology of this literature, the model is not *complete*. It does not specify the

outcomes given the inputs. Figure 1, adapted from CM, shows the different regions in the $(\varepsilon_{Am}, \varepsilon_{Bm})$ space.

The implication of this is that the probability of the outcome $(d_{Am}, d_{Bm}) = (0, 1)$ cannot be written as a function of the parameters of the model, $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, even given distributional assumptions on $(\varepsilon_{Am}, \varepsilon_{Bm})$. Instead the model implies a lower and upper bound on this probability:

$$H_{L,01}(\theta) \leq \Pr\left((d_{Am}, d_{Bm}) = (0, 1)\right) \leq H_{U,01}(\theta).$$

Inspecting Figure 1 shows that

$$\begin{aligned} H_{L,01}(\theta) = &\Pr(\varepsilon_{Am} < -\alpha_A, -\alpha_B < \varepsilon_{Bm}) \\ &+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B - \delta_B < \varepsilon_{Bm}), \end{aligned}$$

and

$$\begin{aligned} H_{U,01}(\theta) = &\Pr(\varepsilon_{Am} < -\alpha_A, \alpha_B < \varepsilon_{Bm}) \\ &+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B - \delta_B < \varepsilon_{Bm}), \\ &+\Pr(-\alpha_A \leq \varepsilon_{Am} < -\alpha_A - \delta_A, -\alpha_B < \varepsilon_{Bm} < -\alpha_B - \delta_B), \end{aligned}$$

Similar expressions can be derived for the probability $\Pr\left((d_{Am}, d_{Bm}) = (1, 0)\right)$. Thus in general we can write the information about the parameters in large samples as

$$\begin{pmatrix} H_{L,00}(\theta) \\ H_{L,01}(\theta) \\ H_{L,10}(\theta) \\ H_{L,11}(\theta) \end{pmatrix} \leq \begin{pmatrix} \Pr\left((d_{Am}, d_{Bm}) = (0, 0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (0, 1)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1, 0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1, 1)\right) \end{pmatrix} \leq \begin{pmatrix} H_{U,00}(\theta) \\ H_{U,01}(\theta) \\ H_{U,11}(\theta) \\ H_{U,11}(\theta) \end{pmatrix}.$$

(For $(d_A, d_B) = (0, 0)$ or $(d_A, d_B) = (1, 1)$ the lower and upper bound coincide, but for ease of exposition we treat all four configurations symmetrically.) The $H_{L,ij}(\theta)$ and $H_{U,ij}(\theta)$ are

known functions of θ. The data allow us to estimate the foru probabilities, which contain only three separate pieces of information because the probabilities add up to one. Given these probabilities, the identified set is the set of all θ that satisfy all eight inequalities. In the simple model above, there are four parameters. Even in the case with the lower and upper bounds for the probabilities coinciding, these would in general not be identified.

We can write this in the GIR form by defining

$$\psi(d_A, d_B|\alpha_A, \alpha_B, \delta_A, \delta_B) = \begin{pmatrix} H_{U,00}(\theta) - (1-d_A)\cdot(1-d_B) \\ (1-d_A)\cdot(1-d_B) - H_{L,00}(\theta) \\ H_{U,01}(\theta) - (1-d_A)\cdot d_B \\ (1-d_A)\cdot d_B - H_{L,01}(\theta) \\ H_{U,10}(\theta) - d_A\cdot(1-d_B) \\ d_A\cdot(1-d_B) - H_{L,10}(\theta) \\ H_{U,11}(\theta) - d_A\cdot d_B \\ d_A\cdot d_B - H_{L,11}(\theta) \end{pmatrix},$$

so that the model implies that at the true values of the parameters

$$\mathbb{E}\left[\psi(d_A, d_B|\alpha_A, \alpha_B, \delta_A, \delta_B)\right] \geq 0.$$

3. Estimation

Chernozhukov, Hong, and Tamer (2007, CHT) consider, among other things, the case with moment inequality conditions,

$$\mathbb{E}[\psi(Z,\theta)] \geq 0,$$

where $\psi(z,\theta)$ is a known vector of functions, of dimension M, and the unknown parameter θ is of dimension K. Let Θ be the parameter space, a subset of $\mathbb{R}^K$.

Define for a vector x the vector $(x)_+$ to be the component-wise non-negative part, and $(x)_-$ to be the component-wise non-positive part, so that for all x, $x = (x)_- + (x)_+$. For a given $M \times M$ non-negative definite weight matrix W, CHT consider the population objective function

$$Q(\theta) = \mathbb{E}[\psi(Z,\theta)]'_- W \mathbb{E}[\psi(Z,\theta)]_-.$$

For all θ in the identified set, denoted by $\Theta_I \subset \Theta$, we have $Q(\theta) = 0$.

The sample equivalent to this population objective function is

$$Q_N(\theta) = \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_-' W \left(\frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta) \right)_- .$$

We cannot simply estimate the identified set as

$$\tilde{\Theta}_I = \{ \theta \in \Theta \,|\, Q_N(\theta) = 0 \} ,$$

The reason is that even for θ in the identified set $Q_N(\theta)$ may be positive with high probability. A simple way to see that is to consider the standard GMM case with equalities and over-identification. If $\mathbb{E}[\psi(Z, \theta)] = 0$, the objective function will not be zero in finite samples in the case with over-identification. As a result, $\tilde{\Theta}_I$ can be empty when Θ_I is not, even in large samples.

This is the reason CHT estimate the set Θ_I as

$$\hat{\Theta}_I = \{ \theta \in \Theta \,|\, Q_N(\theta) \leq a_N \} ,$$

where $a_N \to 0$ at the appropriate rate. In most regular problems $a_N = c/N$, leading to an estimator $\hat{\Theta}_I$ that is consistent for Θ_I, by which we mean that the two sets get close to each other, in the Hausdorf sense that

$$\sup_{\theta \in \Theta_I} \inf_{\theta' \in \hat{\Theta}_I} d(\theta, \theta') \longrightarrow 0, \qquad \text{and} \quad \sup_{\theta' \in \hat{\Theta}_I} \inf_{\theta \in \Theta_I} d(\theta, \theta') \longrightarrow 0,$$

where $d(\theta, \theta') = ((\theta - \theta)'(\theta - \theta'))^{1/2}$.

3. Inference: General Issues

There is a rapidly growing literature concerned with developing methods for inference in partially identified models, including Beresteanu and Molinari (2006), Chernozhukov, Hong, and Tamer (2007), Imbens and Manski (2004), Rosen (2006), and Romano and Shaikh

(2007ab). In many cases the partially identified set itself is difficult to characterize. In the scalar case this can be much simpler. There it often is an interval, $[\theta_{\rm LB}, \theta_{\rm UB}]$. There are by now a number of proposals for constructing confidence sets. They differ in implementation as well as in their goals. One issue is whether one wants a confidence set that includes each element of the identified set with fixed probability, or the entire identified set with that probability. Formally, the first question looks for a confidence set $\mathrm{CI}_\alpha^\theta$ that satisfies

$$\inf_{\theta\in[\theta_{\rm LB},\theta_{\rm UB}]} \Pr\left(\theta \in \mathrm{CI}_\alpha^\theta\right) \geq \alpha.$$

In the second case we look for a set $\mathrm{CI}_\alpha^{[\theta_{\rm LB},\theta_{\rm UB}]}$ such that

$$\Pr\left([\theta_{\rm LB}, \theta_{\rm UB}] \subset \mathrm{CI}_\alpha^\theta\right) \geq \alpha.$$

The second requirement is stronger than the first, and so generally $\mathrm{CI}_\alpha^\theta \subset \mathrm{CI}_\alpha^{[\theta_{\rm LB},\theta_{\rm UB}]}$. Here we follow Imbens and Manski (2004) and Romano and Shaikh (2007a) who focus on the first case. This seems more in line with teh traditional view of confidence interval in that they should cover the true value of the parameter with fixed probability. It is not clear why the fact that the object of interest is not point-identified should change the definition of a confidence interval. CHT and Romano and Shaikh (2007b) focus on the second case.

Next we discuss two specific examples to illustrate some of the issues that can arise, in particular the uniformity of confidence intervals.

3.1 Inference: A Missing Data Problem

Here we continue the missing data example from Section 2.1. We have a random sample of $(W_i, W_i \cdot Y_i)$, for $i = 1, \ldots, N$. Y_i is known to lie in the interval $[0, 1]$, interest is in $\theta = \mathbb{E}[Y]$, and the parameter space is $\Theta = [0, 1]$. Define $\mu_1 = \mathbb{E}[Y|W = 1]$, $\lambda = \mathbb{E}[Y|W = 0]$, $\sigma^2 = \mathbb{V}(Y|W = 1)$, and $p = \mathbb{E}[W]$. For ease of exposition we assume p is known. The identified set is

$$\Theta_I = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)].$$

Imbens and Manski (2004) discuss confidence intervals for this case. The key feature of this problem, and similar ones, is that the lower and upper bounds are well-behaved functionals of the joint distribution of the data that can be estimated at the standard parametric $\sqrt{N}$ rate with an asymptotic normal distribution. In this specific example the lower and upper bound are both functions of a single unknown parameter, the conditional mean μ_1. The first step is a 95% confidence interval for μ_1. Let $N_1 = \sum_i W_i$ and $\overline{Y}_1 = \sum_i W_i \cdot Y_i / N_1$. The standard confidence interval is

$$\text{CI}_\alpha^{\mu_1} = \left[\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}, \overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right].$$

Consider the confidence interval for the lower and upper bound:

$$\text{CI}_\alpha^{p\cdot\mu_1} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) \right],$$

and

$$\text{CI}_\alpha^{p\cdot\mu_1+(1-p)} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right) + (1-p), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

A simple and valid confidence interval can be based on the lower confidence bound on the lower bound and the upper confidence bound on the upper bound:

$$\text{CI}_\alpha^{\theta} = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

This is generally conservative. For each θ in the interior of Θ_I, the asymptotic coverage rate is 1. For $\theta \in \{\theta_{\text{LB}}, \theta_{\text{UB}}\}$, the coverage rate is $\alpha + (1-\alpha)/2$.

The interval can be modified to give asymptotic coverage equal to α by changing the quantiles used in the confidence interval construction, essentially using one-sided critical values,

$$\text{CI}_\alpha^{\theta} = \left[p \cdot \left(\overline{Y} - 1.645 \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + 1.645 \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right].$$

This has the problem that if $p = 0$ (when θ is point-identified), the coverage is only $\alpha-(1-\alpha)$. In fact, for values of p close to zero, the confidence interval would be shorter than the confidence interval in the point-identified case. Imbens and Manski (2004) suggest modifying the confidence interval to

$$\text{CI}^{\theta}_{\alpha} = \left[p \cdot \left(\overline{Y} - C_N \cdot \sigma/\sqrt{N_1} \right), p \cdot \left(\overline{Y} + C_N \cdot \sigma/\sqrt{N_1} \right) + 1 - p \right],$$

where the critical value C_N satisfies

$$\Phi\left(C_N + \sqrt{N} \cdot \frac{1-p}{\sigma/\sqrt{p}} \right) - \Phi\left(-C_N\right) = \alpha.$$

and $C_N = 1.96$ if $p = 0$. This confidence interval has asymptotic coverage 0.95, uniformly over p.

3.2. Inference: Multiple Inequalities

Here we look at inference in the Genereralized Inequality (GIR) setting. The example is a simplified version of the moment inequality type of problems discussed in CHT, Romano and Shaikh (2007ab), Pakes, Porter, Ho, and Ishii (2006), Andrews and Guggenberger (2007), and Hirano and Porter, (2011). Suppose we have two moment inequalities,

$$\mathbb{E}[X] \geq \theta, \qquad \text{and} \quad \mathbb{E}[Y] \geq \theta.$$

The parameter space is $\Theta = [0, \infty)$. Let $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$. We have a random sample of size N of the pairs (X, Y). The identified set is

$$\Theta_I = [0, \min(\mu_X, \mu_Y)].$$

The key difference with the previous example is that the upper bound is no longer a smooth, well-behaved functional of the joint distribution. In the simple two-inequality example, if μ_X is close to μ_Y, the distribution of the estimator for the upper bound is not well approximated by a normal distribution. Suppose we estimate the means of X and Y by

$\overline{X}$, and $\overline{Y}$, and that the variances of X and Y are known to be equal to σ^2. A naive 95% confidence interval would be

$$C_\alpha^\theta = [0, \min(\overline{X}, \overline{Y}) + 1.645 \cdot \sigma/N].$$

This confidence interval essentially ignores the moment inequality that is not binding in the sample. It has asymptotic 95% coverage for all values of μ_X, μ_Y, as long as $\min(\mu_X, \mu_Y) > 0$, and $\mu_X \neq \mu_Y$. The first condition ($\min(\mu_X, \mu_Y) > 0$) is the same as the condition in the Imbens-Manski example. It can be dealt with in the same way by adjusting the critical value slightly based on an initial estimate of the width of the identified set.

The second condition raises a different uniformity concern. The naive confidence interval essentially assumes that the researcher knows which moment conditions are binding. This is true in large samples, unless there is a tie. However, in finite samples ignoring uncertainty regarding the set of binding moment inequalities may lead to a poor approximation, especially if there are many inequalities. One possibility is to construct conservative confidence intervals (e.g., Pakes, Porter, Ho, and Ishii, 2007). However, such intervals can be unnecessarily conservative if there are moment inequalities that are far from binding.

One would like construct confidence intervals that asymptotically ignore irrelevant inequalities, and at the same time are valid uniformly over the parameter space. Bootstrapping is unlikely to work in this setting. One way of obtaining confidence intervals that are uniformly valid is based on subsampling. See Romano and Shaikh (2007a), and Andrews and Guggenberger (2007). Little is known about finite sample properties in realistic settings.

References

Andrews, D., S. Berry, and P. Jia (2004), "Confidence Regions for Parameters in Discrete Games with Multiple Equilibria, with an Application to Discount Chain Store Location,"unpublished manuscript, Department of Economics, Yale University.

Andrews, D., and P. Guggenberger (2004), "The Limit of Finite Sample Size and a Problem with Subsampling,"unpublished manuscript, Department of Economics, Yale University.

Aradillas-Lopez, A., and E. Tamer (2007), "The Identification Power of Equilibrium in Games,"unpublished manuscript, Department of Economics, Princeton University.

Balke, A., and J. Pearl, (1997), "Bounds on Treatment Effects from Studies with Imperfect Compliance,"*Journal of the American Statistical Association*, 92: 1172-1176.

Beresteanu, A., and F. Molinari, (2006), "Asymptotic Properties for a Class of Partially Identified Models,"Unpublished Manuscript, Department of Economics, Cornell University.

Blundell, R., M. Browning, and I. Crawford, (2007), "Best Nonparametric Bounds on Demand Responses,"Cemmap working paper CWP12/05, Department of Economics, University College London.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir, (2007), "Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds,"*Econometrica*, 75(2): 323-363.

Chernozhukov, V., H. Hong, and E. Tamer (2007), "Estimation and Confidence Regions for Parameter Sets in Econometric Models,"forthcoming, *Econometrica*.

Ciliberto, F., and E. Tamer (2004), "Market Structure and Multiple Equilibria in Airline Markets,"Unpublished Manuscript.

Haile, P., and E. Tamer (2003), "Inference with an Incomplete Model of English Auctions,"*Journal of Political Economy*, Vol 111(1), 1-51.

Heckman, J., (1978), "Dummy Endogenous Variables in a Simultaneous Equations

System", *Econometrica*, Vol. 46, 931–61.

Heckman, J. J. (1990), "Varieties of Selection Bias," *American Economic Review* 80, 313-318.

Hirano, K., and J. Porter, (2011), "Impossibility Results for Nondifferentiable Functionals," unpublished manuscript, Department of Economics, University of Arizona.

Honoré, B., and E. Tamer (2006), "Bounds on Parameters in Dynamic Discrete Choice Models," *Econometrica*, 74(3): 611-629.

Hotz, J., C. Mullin, and S. Sanders, (1997), "Boudning Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64(4), 575-603.

Imbens, G., and C. Manski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 74(6): 1845-1857.

Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

Manski, C. (1995), *Identification Problems in the Social Sciences*, Cambridge, Harvard University Press.

Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.

Manski, C., and J. Pepper, (2000), "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(): 997-1010.

Manski, C., G. Sandefur, S. McLanahan, and D. Powers (1992), "Alternative Estimates of the Effect of Family Structure During Adolescence on High School," *Journal of the American Statistical Association*, 87(417):25-37.

Milgrom, P, and R. Weber (1982), "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(3): 1089-1122.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006), "Moment Inequalities and Their Application,"Unpublished Manuscript.

ROBINS, J., (1989), "The Analysis of Randomized and Non-randomized AIDS Trials Using a New Approach to Causal Inference in Longitudinal Studies,"in *Health Service Research Methodology: A Focus on AIDS*, (Sechrest, Freeman, and Mulley eds), US Public Health Service, 113-159.

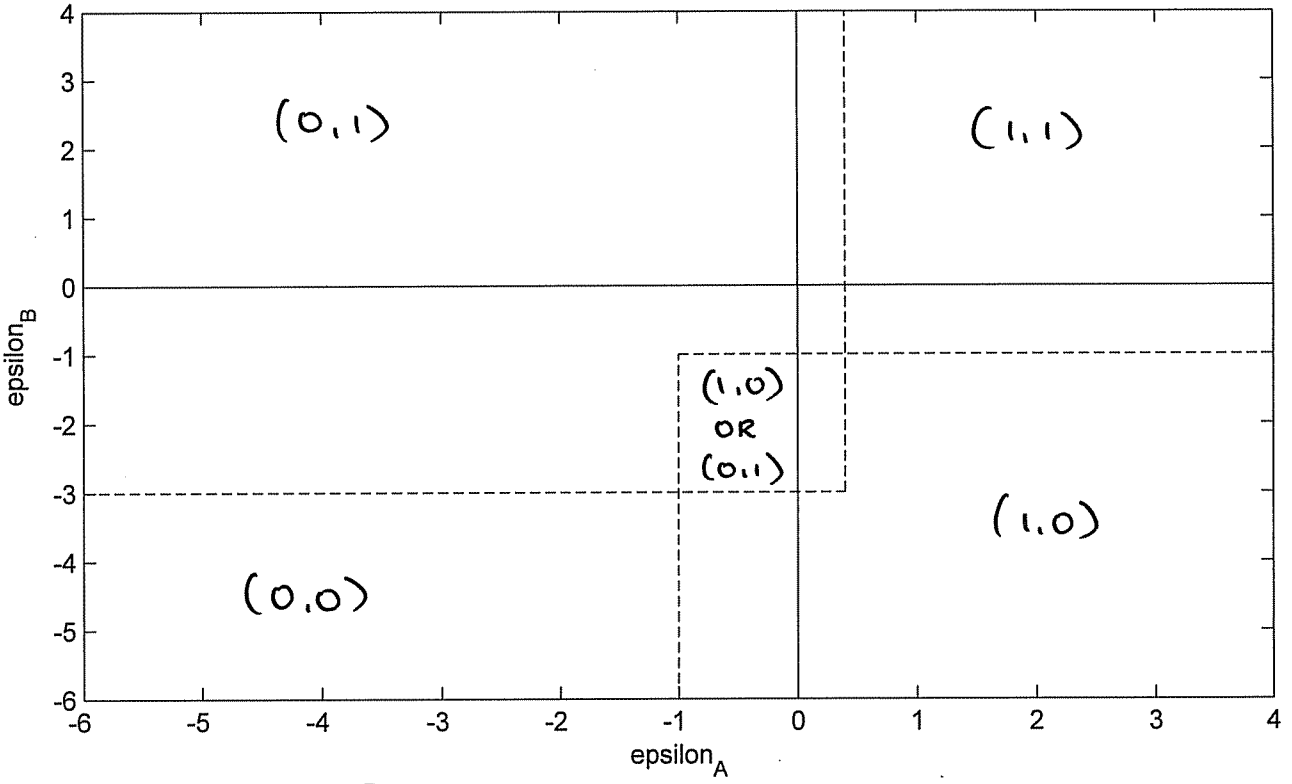
ROMANO, J., AND A. SHAIKH (2006a), "Inference for Partially Identified Parameters,"Unpublished Manuscript, Stanford University.

ROMANO, J., AND A. SHAIKH (2006b), "Inference for Partially Identified Sets,"Unpublished Manuscript, Stanford University.

ROSEN, A., (2005), "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities," Unpublished Manuscript, Department of Economics, University College London.

WOOLDRIDGE, J (2002), "Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity,"*Journal of Applied Econometrics*, 20, 39-54.

Figure 1 (d_A, d_B)



$\alpha_A = 1$ $\delta_A = -1.4$

$\alpha_B = 3$ $\delta_B = -2$

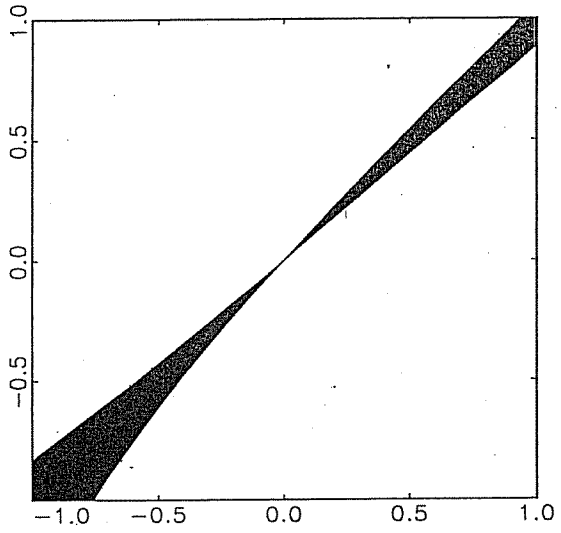


FIGURE 1.—Identified region for γ as a function of its true value.

"Cross-Section Econometrics"

Lecture 5

Partial Identification

Guido Imbens

AEA Lectures, Chicago, January 2012

Outline

1. Introduction
2. Example I: Missing Data
3. Example II: Returns to Schooling
4. Example III: Initial Conditions Problems in Panel Data
5. Example IV: Auction Data
6. Example V: Entry Models
7. Estimation and Inference

1. Introduction

Traditionally in constructing statistical or econometric models researchers look for models that are *(point-)identified*: given a large (infinite) data set, one can infer without uncertainty what the values are of the objects of interest.

It would appear that a model where we cannot learn the parameter values even in infinitely large samples would not be very useful.

However, it turns out that even in cases where we cannot learn the value of the estimand *exactly* in large samples, in many cases we can still learn a fair amount, even in finite samples. A research agenda initiated by Manski has taken this perspective.

Here we discuss a number of examples to show how this approach can lead to interesting answers in settings where previously were viewed as intractable.

We also discuss some results on inference.

1. Are we interested in confidence sets for parameters or for identified sets?
2. Concern about uniformity of inferences (confidence cant be better in partially identified case than in point-identified case).

2. I: Missing Data

If $D_i = 1$, we observe Y_i, and if $D_i = 0$ we do not observe Y_i. We always observe the missing data indicator D_i. We assume the quantity of interest is the population mean $\theta = \mathbb{E}[Y_i]$.

In large samples we can learn $p = \mathbb{E}[D_i]$ and $\mu_1 = \mathbb{E}[Y_i|D_i = 1]$, but nothing about $\mu_0 = \mathbb{E}[Y_i|D_i = 0]$. We can write:

$$\theta = p \cdot \mu_1 + (1 - p) \cdot \mu_0.$$

Since even in large samples we learn nothing about μ_0, it follows that without additional information there is no limit on the range of possible values for θ.

Even if p is very close to 1, the small probability that $D_i = 0$ combined with the possibility that μ_0 is very large or very small allows for a wide range of values for θ.

Now suppose we know that the variable of interest is binary: $Y_i \in \{0, 1\}$. Then natural (not data-informed) lower and upper bounds for μ_0 are 0 and 1 respectively. This implies bounds on θ:

$$\theta \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = [p \cdot \mu_1, p \cdot \mu_1 + (1 - p)] .$$

These bounds are *sharp*, in the sense that without additional information we can not improve on them.

Formally, for all values θ in $[\theta_{\text{LB}}, \theta_{\text{UB}}]$, we can find a joint distribution of (Y_i, W_i) that is consistent with the joint distribution of the observed data and with θ.

We can also obtain informative bounds if we modify the object of interest a little bit.

Suppose we are interested in the median of Y_i, $\theta_{0.5} = \text{med}(Y_i)$.

Define $q_\tau(Y_i)$ to be the τ quantile of the conditional distribution of Y_i given $D_i = 1$. Then the median cannot be larger than $q_{1/(2p)}(Y_i)$ because even if all the missing values were large, we know that at least $p \cdot (1/(2p)) = 1/2$ of the units have a value less than or equal to $q_{1/(2p)}(Y_i)$.

Then, if $p > 1/2$, we can infer that the median must satisfy

$$\theta_{0.5} \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(2p-1)/(2p)}(Y_i), q_{1/(2p)}(Y_i)\right] ,$$

and we end up with a well defined, and, depending on the data, more or less informative identified interval for the median.

If fewer than 50% of the values are observed, or $p < 1/2$, then we cannot learn anything about the median of Y_i without additional information (for example, a bound on the values of Y_i), and the interval is $(-\infty, \infty)$.

More generally, we can obtain bounds on the τ quantile of the distribution of Y_i, equal to

$$\theta_\tau \in [\theta_{\text{LB}}, \theta_{\text{UB}}] = \left[q_{(\tau-(1-p))/p}(Y_i|D_i = 1), q_{\tau/p}(Y_i|D_i = 1)\right] .$$

which is bounded if the probability of Y_i being missing is less than $\min(\tau, 1 - \tau)$.

3. Example II: Returns to Schooling

Manski-Pepper are interested in estimating returns to schooling. They start with an individual level response function $Y_i(w)$.

$$\Delta(s,t) = \mathbb{E}[Y_i(t) - Y_i(s)],$$

is the difference in average outcomes (log earnings) given t rather than s years of schooling. Values of $\Delta(s,t)$ are the object of interest.

W_i is the actual years of school, and $Y_i = Y_i(W_i)$ be the actual log earnings.

If one makes an unconfoundedness/exogeneity assumption that

$$Y_i(w) \perp W_i \mid X_i,$$

for some set of covariates, one can estimate $\Delta(s,t)$ consistently given some support conditions. MP relax this assumption.

Alternative Assumptions considered by MP

Increasing education does not lower earnings:

Assumption 1 (Monotone Treatment Response)
If $w' \geq w$, then $Y_i(w') \geq Y_i(w)$.

On average, individuals who choose higher levels of education would have higher earnings at each level of education than individuals who choose lower levels of education.

Assumption 2 (Monotone Treatment Selection)
If $w'' \geq w'$, then for all w, $\mathbb{E}[Y_i(w)|W_i = w''] \geq \mathbb{E}[Y_i(w)|W_i = w']$.

Under these two assumptions, bound on $\mathbb{E}[Y_i(w)]$ and $\Delta(s,t)$:

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \geq w) + \sum_{v<w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v)$$

$$\leq \mathbb{E}[Y_i(w)] \leq$$

$$\mathbb{E}[Y_i|W_i = w] \cdot \Pr(W_i \leq w) + \sum_{v>w} \mathbb{E}[Y_i|W_i = v] \cdot \Pr(W_i = v).$$

Using NLS data MP estimate the upper bound on the the returns to four years of college, $\Delta(12,16)$ to be 0.397.

Translated into average yearly returns this gives us 0.099, which is in fact lower than some estimates that have been reported in the literature.

This analysis suggests that the upper bound is in this case reasonably informative, given a remarkably weaker set of assumptions.

4. Example III: Initial Conditions Problems in Panel Data (Honoré and Tamer)

$$Y_{it} = 1\{X_{it}'\beta + Y_{it-1} \cdot \gamma + \alpha_i + \epsilon_{it} \geq 0\},$$

with the ϵ_{it} independent $\mathcal{N}(0,1)$ over time and individuals. Focus on γ.

Suppose we also postulate a parametric model for the random effects α_i:

$$\alpha | X_{i1}, \ldots, X_{iT} \sim G(\alpha|\theta)$$

Then the model is almost complete.

All that is missing is:

$$p(Y_{i1}|\alpha_i, X_{i1}, \ldots, X_{iT}).$$

HT assume a discrete distribution for α, with a finite and known set of support points. They fix the support to be $-3, -2.8, \ldots, 2.8, 3$, with unknown probabilities.

In the case with $T = 3$ they find that the range of values for γ consistent with the data generating process (the identified set) is very narrow.

If γ is in fact equal to zero, the width of the set is zero. If the true value is $\gamma = 1$, then the width of the interval is approximately 0.1. (It is largest for γ close to, but not equal to, -1.) See Figure 1, taken from HT.

The HT analysis shows nicely the power of the partial identification approach: A problem that had been viewed as essentially intractable, with many non-identification results, was shown to admit potentially precise inferences. Point identification is not a big issue here.

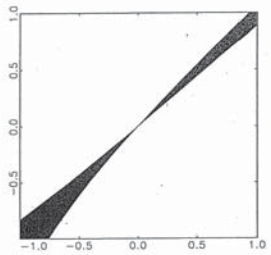


FIGURE 1.—Identified region for γ as a function of its true value.

5. Example IV: Auction Data

Haile and Tamer study English or oral ascending bid auctions. In such auctions bidders offer increasingly higher prices until only one bidder remains. HT focus on a symmetric independent private values model. In auction t, bidder i has a value ν_{it}, drawn independently from the value for bidder j, with cdf $F_\nu(v)$

HT are interested in the value distribution $F_\nu(v)$. This is assumed to be the same in each auction (after adjusting for observable auction characteristics).

One can imagine observing exactly when each bidder leaves the auction, thus directly observing their valuations. This is not what is typically observed. For each bidder we do not know at any point in time whether they are still participating unless they subsequently make a higher bid.

Haile-Tamer Assumptions

Assumption 3 *No bidder ever bids more than their valuation*

Assumption 4 *No bidder will walk away and let another bidder win the auction if the winning bid is lower than their own valuation*

Upper Bound on Value Distribution

Let the highest bid for participant i in auction t be b_{it}. We ignore variation in number of bidders per auction, and presence of covariates.

Let $F_b(b) = \Pr(b_{it} \leq b)$ be the distribution function of the bids (ignoring variation in the number of bidders by auction). This distribution can be estimated because the bids are observed.

Because no bidder ever bids more than their value, it follows that $b_{it} \leq \nu_{it}$. Hence, without additional assumptions,

$$F_\nu(v) \leq F_b(v), \quad \text{for all } v.$$

Lower Bound on Value Distribution

The second highest of the values among the n participants in auction t must be less than or equal to the winning bid. This follows from the assumption that no participant will let someone else win with a bid below their valuation.

Let $F_{\nu,m:n}(v)$ denote the mth order statistic in a random sample of size n from the value distribution, and let $F_{B,n:n}(b)$ denote the distribution of the winning bid in auctions with n participants. Then

$$F_{B,n:n}(v) \leq F_{\nu,n-1:n}(v).$$

The distribution of the any order statistic is monotonically related to the distribution of the parent distribution, and so a lower bound on $F_{\nu,n-1:n}(v)$ implies a lower bound on $F_\nu(v)$.

6. Example V: Entry Models (Cilberto & Tamer)

Suppose two firms, A and B, contest a set of markets. In market m, $m = 1, \ldots, M$, the profits for firms A and B are

$$\pi_{Am} = \alpha_A + \delta_A \cdot d_{Bm} + \varepsilon_{Am},\ \pi_{Bm} = \alpha_B + \delta_B \cdot d_{Am} + \varepsilon_{Bm}.$$

where $d_{Fm} = 1$ if firm F is present in market m, for $F \in \{A, B\}$, and zero otherwise.

Decisions assuming complete information satisfy Nash equilibrium condition

$$d_{Am} = 1\{\pi_{Am} \geq 0\}, \quad d_{Bm} = 1\{\pi_{Bm} \geq 0\}.$$

Incomplete Model

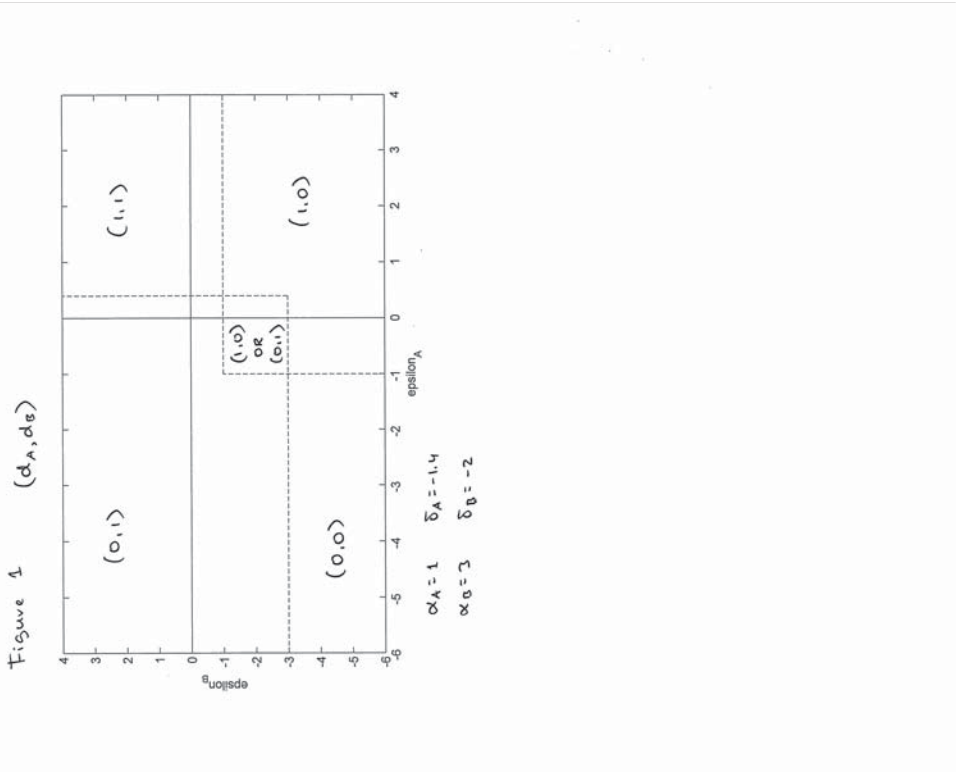
For pairs of values $(\varepsilon_{Am}, \varepsilon_{Bm})$ such that

$$-\alpha_A < \varepsilon_A \leq -\alpha_A - \delta_A, \quad -\alpha_B < \varepsilon_B \leq -\alpha_B - \delta_B,$$

both $(d_A, d_B) = (0, 1)$ and $(d_A, d_B) = (1, 0)$ satisfy the profit maximization condition.

In the terminology of this literature, the model is *incomplete*. It does not specify the outcomes given the inputs. Missing is an equilibrium selection mechanism, which is typically difficult to justify.

Figure 1, adapted from CM, shows the different regions in the $(\varepsilon_{Am}, \varepsilon_{Bm})$ space.



Implication: Inequality Conditions

The implication of this is that the probability of the outcome $(d_{Am}, d_{Bm}) = (0,1)$ cannot be written as a function of the parameters of the model, $\theta = (\alpha_A, \delta_A, \alpha_B, \delta_B)$, even given distributional assumptions on $(\varepsilon_{Am}, \varepsilon_{Bm})$.

Instead the model implies a lower and upper bound on this probability:

$$H_{L,01}(\theta) \leq \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \leq H_{U,01}(\theta).$$

Thus in general we can write the information about the parameters in large samples as

$$\begin{pmatrix} H_{L,00}(\theta) \\ H_{L,01}(\theta) \\ H_{L,10}(\theta) \\ H_{L,11}(\theta) \end{pmatrix} \leq \begin{pmatrix} \Pr\left((d_{Am}, d_{Bm}) = (0,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (0,1)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,0)\right) \\ \Pr\left((d_{Am}, d_{Bm}) = (1,1)\right) \end{pmatrix} \leq \begin{pmatrix} H_{U,00}(\theta) \\ H_{U,01}(\theta) \\ H_{U,11}(\theta) \\ H_{U,11}(\theta) \end{pmatrix}.$$

7.A Estimation

Chernozhukov, Hong, and Tamer study Generalized Inequality Restriction (GIR) setting:

$$\mathbb{E}[\psi(Z, \theta)] \geq 0,$$

where $\psi(z, \theta)$ is known. Fits CT entry example

Define for a vector x the vector $(x)_+$ to be the component-wise non-negative part, and $(x)_-$ to be the component-wise non-positive part, so that for all x, $x = (x)_- + (x)_+$.

For a given $M \times M$ non-negative definite weight matrix W, CHT consider the population objective function

$$Q(\theta) = \mathbb{E}[\psi(Z, \theta)]'_- W \mathbb{E}[\psi(Z, \theta)]_-.$$

For all $\theta \in \Theta_I$, we have $Q(\theta) = 0$, and for $\theta \notin \Theta_I$, we have $Q(\theta) > 0$

The sample equivalent to this population objective function is

$$Q_N(\theta) = \left(\frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \theta)\right)'_- W \left(\frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \theta)\right)_-.$$

We cannot simply estimate the identified set as

$$\tilde{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) = 0\},$$

The reason is that even for θ in the identified set $Q_N(\theta)$ may be positive with high probability, and $\tilde{\Theta}_I$ can be empty when Θ_I is not, even in large samples.

A simple way to see that is to consider the standard GMM case with equalities and over-identification. If $\mathbb{E}[\psi(Z,\theta)] = 0$, the objective function will not be zero in finite samples in the case with over-identification.

This is the reason CHT suggest estimating the set Θ_I as

$$\hat{\Theta}_I = \{\theta \in \Theta \,|Q_N(\theta) \leq a_N\},$$

where $a_N \to 0$ at the appropriate rate.

7.B Inference

Fast growing literature, Beresteanu and Molinari (2006), Chernozhukov, Hong, and Tamer (2007), Galichon and Henry (2006), Imbens and Manski (2004), Rosen (2006), and Romano and Shaikh (2007ab).

First issue: do we want a confidence set that includes each element of the identified set with fixed probability, or the entire identified set with that probability. First

$$\inf_{\theta \in [\theta_{\text{LB}}, \theta_{\text{UB}}]} \Pr\left(\theta \in \text{CI}_\alpha^\theta\right) \geq \alpha.$$

Second

$$\Pr\left([\theta_{\text{LB}}, \theta_{\text{UB}}] \subset \text{CI}_\alpha^{[\theta_{\text{LB}}, \theta_{\text{UB}}]}\right) \geq \alpha.$$

The second requirement is stronger than the first, and so generally $\text{CI}_\alpha^\theta \subset \text{CI}_\alpha^{[\theta_{\text{LB}}, \theta_{\text{UB}}]}$.

7.B.I Well behaved Estimators for Bounds

Missing data example, (p, prob of missing data, known). Identified set:

$$\Theta_I = [p \cdot \mu_1, p \cdot \mu_1 + (1-p)].$$

Standard interval for μ_1:

$$\text{CI}_\alpha^{\mu_1} = \left[\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}, \overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1}\right].$$

Three ways to construct 95% confidence intervals for θ.

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.96 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.96 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This is conservative. For each θ in the interior of Θ_I, the cov rate is 1. For $\theta \in \{\theta_{\text{LB}}, \theta_{\text{UB}}\}$, if $p < 1$, the cov rate is 0.975.

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - 1.645 \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + 1.645 \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right].$$

This has the problem that if $p = 1$ (when θ is point-identified), the coverage is only 0.90. Imbens and Manski (2004) suggest modifying the confidence interval to

$$\text{CI}_\alpha^\theta = \left[p \cdot \left(\overline{Y} - C_N \cdot \sigma/\sqrt{N_1}\right), p \cdot \left(\overline{Y} + C_N \cdot \sigma/\sqrt{N_1}\right) + 1 - p\right],$$

where the critical value C_N satisfies

$$\Phi\left(C_N + \sqrt{N} \cdot \frac{1-p}{\sigma/\sqrt{p}}\right) - \Phi\left(-C_N\right) = 0.95$$

This confidence interval has asymptotic coverage 0.95, uniformly over p, for $p \in [p_0, 1]$.

7.B.II Irregular Estimators for Bounds

Simple example of Generalized Inequality Restrictions (GIR) set up.

$$\mathbb{E}[X] \geq \theta, \quad \text{and} \quad \mathbb{E}[Y] \geq \theta.$$

The parameter space is $\Theta = [0, \infty)$. Let $\mu_X = \mathbb{E}[X]$, and $\mu_Y = \mathbb{E}[Y]$. We have a random sample of size N of the pairs (X, Y). The identified set is

$$\Theta_I = [0, \min(\mu_X, \mu_Y)].$$

A naive 95% confidence interval would be

$$C_\alpha^\theta = [0, \min(\overline{X}, \overline{Y}) + 1.645 \cdot \sigma/N].$$

This confidence interval essentially ignores the moment inequality that is not binding in the sample. It has pointwise asymptotic 95% coverage for all values of μ_X, μ_Y, as long as $\min(\mu_X, \mu_Y) > 0$, and $\mu_X \neq \mu_Y$.

The first condition ($\min(\mu_X, \mu_Y) > 0$) is the same as the condition in the Imbens-Manski example. It can be dealt with in the same way by adjusting the critical value slightly based on an initial estimate of the width of the identified set.

The naive confidence interval essentially assumes that the researcher knows which moment conditions are binding. This is true in large samples, unless there is a tie.

However, in finite samples ignoring uncertainty regarding the set of binding moment inequalities may lead to a poor approximation, especially if there are many inequalities. One possibility is to construct conservative confidence intervals (e.g., Pakes, Porter, Ho, and Ishii, 2007). However, such intervals can be unnecessarily conservative if there are moment inequalities that are far from binding.

One would like construct confidence intervals that asymptotically ignore irrelevant inequalities, and at the same time are valid uniformly over the parameter space. Subsampling (but not bootstrapping) appears to work theoretically. See Romano and Shaikh (2007a), and Andrews and Guggenberger (2007). Little is known about finite sample properties in realistic settings.