Specification Testing with Prediction Criterion: Causality, Prediction, and External Validity

Masahiro Kato, Cyberagent, Inc.

Introduction

Model specification testing.

- Regression analysis relies on the correctness of model specification.

e.g., Durbin–Wu–Hausman test

- Correct model: orthogonality of the dependent variables and error term.

Prediction-based model specification test

- Assume availability of train data $\{(X_i, Y_i)\}_{i=1}^n$ and test data $\{\tilde{X}_j\}_{j=1}^m$.
- **Test data**: the data that we want to predict the outcome.
- There are only covariates $\tilde{X}_j$, and the target variables are unobservable.

New definition of correct models.

- Idea: If the model can predict target variables well, the model is correct.
- Under the definition, we show
- The asymptotic distribution of the least squares under covariate shift.
- The asymptotic distribution of the test statistics.

1. Covariate Shift Problem

Data-generating process (DGP):

- There are two stratified data:

$$(X_i, Y_i) \sim p(x, y), \qquad (\tilde{X}_j, \tilde{Y}_j) \sim q(x, y),$$

where $X_i, X_j \in \mathbb{R}^d$ and $Y_i, \tilde{Y}_j \in \mathbb{R}$. $\tilde{Y}_j$ is unobservable.

- Observations:

$$\{(X_i, Y_i)\}_{i=1}^n \sim p(x, y), \qquad \{\tilde{X}_j\}_{j=1}^m \sim q(x),$$

- Furthermore, we put the following assumption on the conditional pdf:

$$p(x, y) = p(y|x)p(x),$$
$$q(x, y) = p(y|x)q(x).$$

- $p(y|x)$ **is invariant across the two data.**
- $p(x)$ and $q(x)$ can be changed
- $p(x)$ and $q(x)$ have a common support.

This setting is called learning under covariate shift.

2. Definition of Correct Model

Linear model:

- Assume a linear model of $\mathbb{E}[Y_i|X_i]$ as $\boldsymbol{Z}^\top(\boldsymbol{X_i})\boldsymbol{\beta}^*$.
- $Z(\cdot)$ is a mapping from X_i to some linear models.

Definition of correct model

- Our model specification is defined from the viewpoint of prediction.
- Parameter that minimizes the MSE over $p(x, y)$ is defined as

$$\alpha_0 = \mathrm{argmin}_b \mathbb{E}_{p(x,y)}[(Y_i - Z^\top(X_i)b)^2].$$

- Parameter that minimizes the MSE over $q(x, y)$ is defined as

$$\gamma_0 = \mathrm{argmin}_b \mathbb{E}_{q(x,y)}\left[\left(\tilde{Y}_j - Z^\top(\tilde{X}_j)b\right)^2\right].$$

- If $\boldsymbol{\alpha_0 = \gamma_0}$, the model is specified correctly
- If $\boldsymbol{\alpha_0 \neq \gamma_0}$, the model is misspecified.
- By using this definition, consider the following hypothesis:

$$\boldsymbol{\mathcal{H}_0: \alpha_0 = \gamma_0 \text{ and } \mathcal{H}_1: \alpha_0 \neq \gamma_0}$$

- If $\mathcal{H}_0$ is rejected, the model specification is incorrect.

3. Covariate Shift Adaptation

- However, we cannot observe $\tilde{Y}_i$.
- Let us define a parameter estimated from $\{(X_i, Y_i)\}_{i=1}^n$ as

$$\boldsymbol{\hat{\alpha} = argmin_b \widehat{\mathbb{E}}_{p(x,y)}[(Y_i - Z^\top(X_i)b)^2]},$$

where $\widehat{\mathbb{E}}_{p(x,y)}$ denotes the sample average of the samples from $p(x, y)$.

- Then, for $\{\tilde{X}_j\}_{j=1}^m$, we define the following estimator:

$$\boldsymbol{\hat{\gamma} = argmin_b \widehat{\mathbb{E}}_{q(x,y)}\left[\left(\tilde{Y}_j - Z^\top(\tilde{X}_j)b\right)^2\right]}$$

$$\boldsymbol{\approx argmin_b \widehat{\mathbb{E}}_{p(x,y)}\left[(Y_i - Z^\top(X_i)b)^2 \frac{q(X_i)}{p(X_i)}\right]}.$$

- **Thus, we approximate $\mathbb{E}_{q(x,y)}$ by using $\widehat{\mathbb{E}}_{p(x,y)}$ and $\frac{q(X_i)}{p(X_i)}$.**

- Let us denote the density ratio $\frac{q(x)}{p(x)}$ by $r^*(x)$.
- We can estimate the density ratio with machine learning methods.

e.g., uLSIF (Kanamori et al. (2012)).

4. Double/Debiased Least Squares Estimator

- **Consider the asymptotic distribution of $\hat{\gamma}$.**
- The density ratio is estimated by machine learning methods.

→ The estimator does not satisfy Donsker's condition.

- We use **double/debiased machine learning** to avid this problem.
- An estimator $\hat{\gamma}$ with a doubly robust form.
- Cross-fitting.
- Doubly robust estimator of the MSE over $q(x, y)$:

$$\widehat{\mathbb{E}}_{q(x,y)}\left[\left(\tilde{Y}_j - Z^\top(X_j)b\right)^2\right]$$
$$\approx \widehat{\mathbb{E}}_{p(x,y)}\left[\left((Y_i - Z^\top(X_i)b)^2 - \left(\hat{f}(X_i) - Z^\top(X_i)b\right)\right)\hat{r}(X_i)\right] + \widehat{\mathbb{E}}_{q(x)}\left[\left(\hat{f}(\tilde{X}_j) - Z^\top(\tilde{X}_j)b\right)^2\right],$$

- $\hat{f}(x)$ is some consistent estimator of $f^*(x) = \mathbb{E}[Y_i \mid x]$.
- $\hat{r}(x)$ is some consistent estimator of $r^*(x) = \frac{q(x)}{p(x)}$.
- We construct the empirical MSE by using cross-fitting.
- Then, if $n = m = N$,

$$\boldsymbol{\sqrt{N}(\hat{\gamma} - \gamma^*) = \mathcal{N}(0, \tilde{\Sigma})}$$

5. Hypothesis Testing

- We construct the test statistics to investigate the hypothesis

$$\mathcal{H}_0: \alpha_0 = \gamma_0 \text{ and } \mathcal{H}_1: \alpha_0 \neq \gamma_0$$

- A standard choice is to use Wald statistics.
- We can construct Wald statistics by using the estimators $\hat{\alpha}$ and $\hat{\gamma}$.
- **The Wald statistics follows $\boldsymbol{\chi^2(k)}$ distribution.**
- $\boldsymbol{k}$ **is the dimension of the linear model.**
- We conduct hypothesis testing using the test statistics.
- If the null hypothesis reject, we can say that model is misspecified.

References

Shimodaira, H. (2000), "Improving predictive inference under covariate shift by weighting the log-likelihood function," Journal of Statistical Planning and Inference.

Kanamori, T., Hido, S., and Sugiyama, M. (2009), "A Least-squares Approach to Direct Importance Estimation," Journal of Machine Learning Research.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/debiased machine learning for treatment and structural parameters," Econometrics Journal.