Expanding Access to Administrative Data for Research in the United States

David Card, UC Berkeley
Raj Chetty, Harvard
Martin Feldstein, Harvard
Emmanuel Saez, UC Berkeley

January 2011

MOTIVATION

High quality data are key to empirical research in social sciences

Comprehensive socio-economic micro data are created by governments for the administration of taxes and programs

Such administrative data are vastly superior to traditional survey data

Modern computers can easily process administrative data for research

Key Research Priority: Develop **direct** and **secure** access to US administrative data for research

ADMINISTRATIVE DATA VS. SURVEY DATA

Administrative data vastly superior to survey data (such as CPS or PSID) along 3 key dimensions:

1) Much Larger Sample Size (full population coverage)

2) Natural Longitudinal Structure

3) Much less measurement error, imputations, and attrition issues

Administrative data cover many socio-economic variables (birth, education, work and income, housing, family, health, retirement, death)

ERODING US LEADERSHIP

US led the way in development of

(1) Micro survey data (CPS, PSID)

(2) Micro-econometrics research in social sciences

Frontier empirical research is shifting from survey to administrative data

US lags behind in providing secure access to admin data for research

⇒ Frontier research is shifting away from US data

LESSONS FROM EXISTING EXPERIENCES

US has developed a number of valuable admin data access initiatives: (1) health (CMS and hospitals), (2) K-12 education (states), (3) earning records (LEHD-Census), (4) earnings-retirement (SSA), (5) income and taxes (IRS)

A number of countries (Scandinavia) have developed systematic and secure access to comprehensive administrative data:

Central statistical agency (1) obtains and maintains admin data from all agencies, (2) prepares de-identified data for each approved research project, (3) provides secure access of the data to researchers, (4) only research output is publicly disclosed

KEY CONDITIONS FOR PROGRAM SUCCESS

(1) Fair and open competition based on scientific merit

(2) Sufficient resources to accommodate all worthy projects

(3) Direct and secure access to (de-identified) micro-data remotely or through local offices

(4) Inclusion of students in teams with data access

Direct access to micro-data is critical for success: Synthetic data or sending computer programs from the outside are not good alternatives

PERSPECTIVES FOR THE UNITED STATES

Central agency model (Scandinavia) most efficient in principle. But this model is unlikely to work in the US:

(1) US government is far more decentralized

(2) US agencies are covered by different disclosure statutes

(3) Distrust of centralized and monopoly government control

(4) Many agencies have sophisticated statistical divisions with valuable internal knowledge

⇒ better to build upon the existing decentralized system

VALUE OF COMPETITION

Plurality of US govt agencies (and hence data sources) could be turned into an advantage using forces of competition

(1) Each agency is invited to develop secure data access for scientific research

(2) Each agency is rewarded by major research funders (NSF, NIH, etc.) for performance in scientific production (easily measurable)

(3) This would unleash the forces of innovation as agencies compete for the best research

(4) Agencies could set up partnerships for merging data from two or more agencies

OMB just started an "evaluation initiative" to provide resources to agencies for scientific evaluations