Overview

I will talk about
1. U.S. Data
2. European Register Data

General Comments

For subtle effects or for instrumental variables estimation, probably need lots of observations.

Finding a dataset with plausibly exogenous variation in an X variable can be key – the Y variable may be negotiable!

If you want to do selection on observables (e.g. OLS or matching), important to have a dataset with a very rich set of controls.

Why use U.S. data? It can make it easier to publish in top journals.

United States

ICPSR (www.icpsr.org) is a user supported data repository at the University of Michigan containing thousands of data sets. Most major universities subscribe to ICPSR and data is downloadable from the site after registration. It has data from many countries.

Census Public Use Micro Samples
1 and 5 percent cross-sectional samples of the U.S. in census years.
Has information on earnings, hours, education, age, race, other basic demographics such as immigrant status.

American Community Survey (2000-2007)
1% sample on an annual basis.
www.ipums.org (requires registration but it is free)

Current Population Survey
Monthly labor market survey of about 60K households
http://www.nber.org/data/cps_index.html
The UCD library has/had access to Unicon CPS Utilities which make it very easy to make extracts from the CPS.
It contains data, documentation and Windows software to help researchers find and extract data.

The Historical Series of Annual Supplements includes Computer Usage, Tobacco Usage, Outgoing Rotations, Food Security, Child Support/Alimony, Immunization, Veterans surveys etc., and holds data from as early as 1962.

March Supplement (Annual Demographic Survey) has information on public and private health insurance, demographics, income and employment.
Since 1995, March CPS has self-reported health status.

Respondents in CPS for 4 consecutive months, then out for 8 months, then in again for 4 months. So, possible to create 2-year panels and study changes.

National Longitudinal Surveys of Youth, 1979 and 1997
http://www.bls.gov/nls/

Longitudinal data of young adults

Follows people from about age 14 into their 30s.

Information on education, wages, labour supply, families, fertility, height, weight, alcohol and drug use, general health measures, attitudes etc.

Very low attrition rates.

Often several siblings from the same family so can use family fixed effects models.

Key attribute: presence of aptitude test score – Armed Forces Qualifying Test (AFQT).

Panel Study of Income Dynamics (PSID)
http://psidonline.isr.umich.edu/
Longitudinal data set of 5000 families beginning in 1968.

Long panel dataset with full range of ages and information on lots of subjects. Particularly strong on labour market variables such as wages, earnings, job tenure and family incomes.

Also has information on health, wealth, consumption, marital history, fertility.

Survey was every year, now every 2 years.

A feature is its length – from 1968 to the present. Implies that it is very useful for intergenerational research, dynamic models of earnings.

Health and Retirement Survey (HRS)

Longitudinal survey conducted every two years.

Covers a broad range of topics, including health, income, assets, employment, retirement, insurance, and family structure.

Mostly used for retirement research.

Nationally representative of the older population.

Covers cohorts born 1931-1941. The version to use is the Rand cleaned up version.
http://hrsonline.isr.umich.edu/meta/rand/index.html

National Health Interview Surveys:
http://www.cdc.gov/nchs/nhis.htm

Annual survey of 60k households, designed to measure the stock of health in the U.S.

Collects demographic information on each member from each family in the house and collects data on topics including health status and limitations, injuries, healthcare access and utilization, health insurance, health behaviours, and income and assets.

Topics covered in Supplements are Cancer Screening, Complementary and Alternative Medicine, Children's Mental Health, and Healthcare Utilization.

Survey of Income and Program Participation (SIPP)

The main objective of SIPP is to provide accurate and comprehensive information about the income and program participation of individuals and households in the United States.

SIPP offers detailed information on cash and noncash income. The survey also collects data on taxes, assets, liabilities, and participation in transfer programs.

SIPP panels usually last between 2.5 and 4 years. Survey carried out every four months.

Sample sizes of 10000 – 50000 households.

Also good for health research – has work disability history, disability status of children, medical expenses, health status, health care utilisation.

Possible to merge some SIPPs with social security earnings records

American Time Use Survey (2003-)

Uses random sample of CPS ex-participants.

Monthly survey with about 1000 persons per month.

Respondent describes activities the previous day for the full 24 hour period.

Lists primary activity and when started and finished.

Can see how much time people spend at work, on housework, playing with their children etc.

Also, can see exactly when various activities occur.

European Register Data

- Available in Finland, Sweden, Norway, Denmark, Portugal, Netherlands
- Difficult to access. Best bet is find a local co-author
- Key feature: common individual identifier across datasets.

Example: Norwegian Datasets that can be Linked

Labour Market Register

- Longitudinal dataset on the **entire population** of Norwegians aged 16-74 in the 1986-2006 period
- Information on educational attainment, earnings, and demographics.
- Can link to biological parents and so link to siblings and calculate family size, birth order.
- Can identify twin pairs

Twin Survey for 1967-1979 cohorts

- Contains Zygosity Information

Military Records

- Males approximately 18-20 from 1987-2006
- Information on IQ, height, weight, mental health

- The IQ measure is the mean of three IQ tests -- arithmetic, word similarities, and figures
  - The arithmetic test is similar to that in the Wechsler Adult Intelligence Scale (WAIS)
  - The word test is similar to the vocabulary test in WAIS
  - The figures test is similar to the Raven Progressive Matrix test

Birth Register

- Birth records for all Norwegian births from 1967-2006 with information on birth weight, gestation, age of mother, infant health, twin births.
- APGAR scores included
- Also has 1-year infant mortality.

School Data

- Have information on lower secondary school cohorts
- We know what school each person attended in their final year (9th grade)
- Data from 1975 onwards

Worker-Firm Matched Data

- From 1984
- Plant and firm identifiers matched to workers.
- Information on industry
- Also information on job titles of workers