One of the themes of this course is that good empirical work in public finance requires good data on economic behavior. This note is meant to introduce you to some of the more common major publicly available data sources that have been used to look at empirical tax questions so that if you have ideas you could examine them on your own.

Some of the best places to look for data sources that might interest you are the websites of the BLS, the Census, and the IRS

http://www.bls.gov
http://www.census.gov

But some of the most interesting and most important work in public finance has come from people finding their own data sources. ALWAYS BE ON THE LOOKOUT FOR DATA. Surveys of occupations, trade association data, and private market research companies are often good sources.

The major databases that many people have used include

**STATISTICS OF INCOME**
Source: IRS/Treasury
Years: Aggregate Data 1913-present
       Micro Data 1960-1996 (available at NBER)

The Statistics of Income (SOI) program produces data files compiled from tax and information returns filed with the IRS. Shown to the right are topics covering these files, as well as files from other IRS sources.

It is aggregate data though most of the books will be broken down by cross-tabs such as by income by state or by different size firms, etc.

Since all federal tax revenue goes through the IRS, these books cover every conceivable topic. There is a micro data source on individual returns from 1978 to 1996. Privacy concerns have made it so that these files may not be updated in the future.
CURRENT POPULATION SURVEY
Source: Bureau of Labor Statistics
Years available: 1958-present
http://www.bls.census.gov/cps/cpsmain.htm
(available from the NBER, too)

The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population. Estimates obtained from the CPS include employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, marital status, and educational attainment. They are also available by occupation, industry, and class of worker. Supplemental questions to produce estimates on a variety of topics including school enrollment, computer usage income, previous work experience, health, employee benefits, and work schedules are also often added to the regular CPS questionnaire.

SURVEY OF INCOME AND PROGRAM PARTICIPATION
Source: Census
Years 1984-1996
http://www.sipp.census.gov/sipp/sipphome.htm

The SIPP collects source and amount of income, labor force information, program participation and eligibility data, and general demographic characteristics to measure the effectiveness of existing federal, state, and local programs; to estimate future costs and coverage for government programs, such as food stamps; and to provide improved statistics on the distribution of income in the country. The SIPP content is built around a "core" of labor force, program participation, and income questions designed to measure the economic situation of persons in the United States. These questions expand the data currently available on the distribution of cash and noncash income and are repeated at each interviewing wave.

The survey design is a continuous series of national panels, with sample size ranging from approximately 14,000 to 36,700 interviewed households. The duration of each panel ranges from 2 1/2 years to 4 years. The SIPP sample is a multistage-stratified sample of the U.S. civilian noninstitutionalized population. For the 1984-1993 panels, a new panel of households was introduced each year in February. A new 4-year 1996 panel was introduced in April 1996.

The survey has been designed also to provide a broader context for analysis by adding questions on a variety of topics not covered in the core section. These questions are labelled "topical modules" and are assigned to particular interviewing waves of the survey. Topics covered by the modules include personal history, child care, wealth, program eligibility, child support, disability, school enrollment, taxes, and annual income.
We release the data periodically in cross-sectional, topical module, and longitudinal reports. We also release public use files containing the core data on income recipiency and program participation. These files are available currently for all waves of the 1984 through 1993 panels and Wave 1 of the 1996 panel. Topical module files containing core and topical module data also are available for these panels for 1984 through 1988 and 1990 through Wave 8 of the 1993 panel. Longitudinal files are also available for the 1984-1993 panels as well as for Waves 1-5 of the 1990 panel and for Waves 1-7 of the 1992 panel.

**NATIONAL LONGITUDINAL SURVEY OF YOUTH**
Source: Bureau of Labor Statistics
Years: 1978-present
http://www.bls.gov/nls/

The National Longitudinal Surveys (NLS) are a set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women. The sample size is about 12,000 and follows the same people over time. A new NLSY started in 1997 with about 9000 respondents and continues to the present, as well.

The NLSY includes numerous variables about family background and about sociology-type factors such as crime, etc. in addition to standard labor type variables.

**SURVEY OF CONSUMER FINANCES**
Source: Federal Reserve
Panel 1983-1989

The Survey of Consumer Finances (SCF) is a triennial survey of the balance sheet, pension, income, and other demographic characteristics of U.S. families. The survey also gathers information on the use of financial institutions. The links to the surveys provide summary results of the surveys, codebooks and related documentation, and the publicly available data. Also included here are the data and related information from the 1962 Survey of Financial Characteristics of Consumers and the 1963 Survey of Changes in Family Finances. These surveys are the most direct precursors of the SCF. Similar information on the 1983 and 1989 Surveys of Pension Providers is provided as well. The sample size is about 4500.

**COMPUSTAT**
Source: Standard and Poor's
Years: 1960s-present
http://www.compustat.com

Financial and accounting data on publicly traded companies. Compiled by a private company. Used by much of the work on taxes and firms.
With data on more than 20,000 companies, no other software integrates as much in-depth financial information into a single, easy-to-access resource as Research Insight from Standard & Poor's Compustat. More than 10,100 active companies with up to 20 years and 48 quarters of fundamental data, and 20 years of monthly market history. More than 10,200 inactive (Research) companies with up to 20 years of history. More than 1,000 Canadian companies with up to 20 years of annual data and quarterly history from 1st quarter 1991, and monthly high, low & close prices, dividends, EPS, book value, shares traded, and common stock float shares as far back as 1982 plus 16 indexes. Extensive fundamental data coverage of over 1,400 data items, ratios & growth rates, monthly stock price data (PDE), industry segment (business segment-line of business) and geographic data, multiple SIC codes, and company name and address information.

There is also an international data base with comprehensive data on more than 15,900 companies in 72 countries around the world. Comprehensive company financial data, extensive balance sheet, income statement, and detailed footnotes.

CONSUMER EXPENDITURE SURVEY
Source: Bureau of Labor Statistics
Years: 1980-present
http://www.bls.gov/cex/home.htm
(available from the NBER, too)

This program consists of two surveys—the quarterly Interview survey and the Diary survey—that provide information on the buying habits of American consumers, including data on their expenditures, income, and consumer unit (families and single consumers) characteristics. Annual income and expenditures integrated from the Interview and Diary surveys in varying detail, classified by income, age, consumer unit size, and other demographic characteristics of consumer units, since 1984. Annual income and expenditures from the Interview and Diary surveys by selected consumer unit characteristics, since 1980.

The surveys target the total noninstitutionalized population (urban and rural) of the United States in 1980, 1984 and thereafter. The surveys targeted the urban noninstitutionalized population in 1981 through 1983. The data are collected in independent quarterly Interview and weekly Diary surveys of approximately 7,500 sample households (5,000 prior to 1999). Each survey has its own independent sample, and each collects data on household income and socioeconomic characteristics. The Interview survey includes monthly out-of-pocket expenditures such as housing, apparel, transportation, health care, insurance, and entertainment. The Diary survey includes weekly expenditures of frequently purchased items such as food and beverages, tobacco, personal care products, and nonprescription drugs and supplies.

TAXSIM
Source: NBER

TAXSIM is a program that will take the data on a person such as given in the micro-data from the
IRS or from the CPS and compute a predicted marginal tax rate. It is frequently used in micro-data studies of individuals.

**Tax Rate Data**

Federal tax rates can be found in the appendix to the wonderful book by Joe Pechman, *Federal Tax Policy*. Old editions of the Statistics of Income also tend to have the statutory rates in them. State rates can be found in publications such as *Significant Features of Fiscal Federalism* and *The Book of the States*. State department of revenue pages also tend to have the rates as do the actual IRS tax forms themselves. The statistics program STATA also has an embedded function that will compute the marginal income tax rate for a given year and income level.