Economic Statistics

How to Find, Use, and Understand Data on the Economy

Clopper Almon
Department of Economics
University of Maryland

May 1998

Contents

Introduction

What was the inflation rate last month? How bad is unemployment? How fast - or slowly - is the economy growing? Answers to all of these questions require *statistics*. We will be studying how they are made and how we can learn from them.

The word *statistics* came into English about 1770 from German where it had meant (since about 1747) a study “ascertaining the political strength” of a *state*. In English, however, it acquired a broader meaning expressed by J. Sinclair in 1798 as “an inquiry for the purpose of ascertaining the quantum of happiness enjoyed by [a country’s] inhabitants.” The word “quantum” here is especially telling. By the end of the 19th century, statistics was clearly the study of facts that can be expressed as numbers. It also became clear that mathematical methods developed in astronomy and in the analysis of games of chance had remarkable applicability to the study of the social statistics. Today, the applications of statistics include almost every public policy discussion, the design of political and marketing campaigns, business planning , and scientific research in fields as diverse as medicine and economics. The word *statistics* is now often used to refer to a common core of mathematical methods which are used for the analysis of data in all these areas. While this book will certainly present some of those methods, we shall not loose sight of the broader meaning of statistics.

Unlike the natural world around us, statistics have to be created by human endeavor in collecting and organizing information. In the United States, a complicated structure of institutions and programs have developed to produce statistics broadly related to the economy. Our course begins with a survey of these institutions and the statistics they make and why they make them. We will learn how to find information in a convenient compendium of U.S. statistics, the *Statistical Abstract of the United States*, which is available inexpensively both as a book and, with much more data, as a CD-ROM. You will also learn to use the Internet to update information from the *Statistical Abstract.*.

Having looked at some of the finished products of data collection and organization, we next turn to what might be called primary statistics, data more or less as it comes into the collecting agency. While we will not actually collect statistics in this course, we will do the next best thing; we will look at a release for public use of data from the 1% sample of the 1990 Census of Population and Housing. We will be dealing with the individual responses to the long form of the Census questionnaire. (Some information has been removed to protect the privacy of the respondent and some has been imputed where there were blanks in the responses.) No one can begin to look at and digest all these data, so how do we make them speak and tell their story? We will first learn some ways to describe a single variable, say family income. We will ask what we can say on the basis of a sample about the characteristics of the whole population. Then, using a technique called *regression analysis*, we will look at relations between one variable and one or more other, explanatory variables.. We will apply all of these methods to data from the 1% sample.

Actually, the 1% sample is so large that we will work with twenty sub-samples. Different students will have different samples, so we can see for ourselves that different samples yield different results. Since most statistics are based on samples, that difference raises the question of what we can say about the total population on the basis of the analysis of a sample. That question is the central subject of *statistical inference*.

After this study of primary data, we return to the study of secondary, or “prepared” data, the end product of the statistical agencies. We will look at the conceptual and measurement challenges in making three of the most used components of the statistical system: the national accounts, the consumer price indexes, and measures of money. Finally, we will see that regression analysis, previously developed for studying primary data, can also be applied to the time series which are produced by the statistical agencies. Here, with data drawn from the national accounts and Federal Reserve statistics, we can investigate questions of macroeconomics such as What influence, if any, do interest rates have on investment?

This broad view of statistics, consistent with its original meaning, makes this course radically different from the course usually taught under this name. The usual course concentrates on probability theory and statistical inference with little or no attention to the other matters which will also concern us. The usual course also works with small amounts of data that is often fictional. We will work realistic volumes of real data. I believe that what you really need to know about probability and statistical inference you will learn here in a way that will make it meaningful to you.

All of the computations for this course can be performed with any one of the three common spreadsheet programs: Lotus 1-2-3, Corel’s Quattro Pro, and Microsoft’s Excel. I have tried all three and helped students in using all three. My preference is for Lotus for what we have to do here. Instructions given here therefore apply to specifically to Lotus, but most of them also apply fairly well to the others. I should make clear that all three are extremely limited as statistical programs. For going beyond this course, you would certainly want to use a different program. The spreadsheets are, however, very visual and virtually universally available; those features makes them attractive for beginning work with statistics.

As to mathematical background, I will assume that you have had about one semester of calculus and are familiar with the elements of both differentiation and integration. In explaining regression, it is convenient to use matrix notation, which will be explained here.

Chapter 1

The Statistical System of the United States

The *Statistical Abstract of the United States*, published by the U.S. Department of Commerce, Bureau of the Census, is both a guide to the U.S. statistical system and a convenient compendium of the major results of the efforts of the Census Bureau and other organizations, both public and private, both domestic and international. We will use the contents of this comprehensive book to illustrate what is found in the different statistical sources. The *Abstract* now appears both as a book and as a CD-ROM. The CD is much the easier way to use it because of the excellent indexing, instant movement from index to table, and the automatic access to more detailed and complete tables in the form of worksheets for Lotus 1-2-3. It even has automatic Internet links for some of the tables which allow easy updating of series or finding more complete material. All of the questions posed in the following discussions are to be answered by using the material from the CD. If you have the a personal computer with a CD-ROM drive running Windows and have purchased the CD, you can answer them all at home. Otherwise, you can use your university's computer laboratory. (Details appear at the end of this chapter.) For further work in this course, you will need Lotus 1-2-3; you can get the same results with Corel Quattro Pro or Microsoft Excel if you are prepared to do a bit of technical work with your system setup. The instructions in this text will be for 1-2-3.

Some seventy federal agencies gather and publish statistics of interest to the general public. For most of them, production of statistics are incidental to their main work. For at least six of them, however, the production of statistics is the central concern. The main providers of economic statistics are the U.S. Bureau of the Census, the Bureau of Labor Statistics, the Bureau of Economic Analysis, the Board of Governors of the Federal Reserve System, the Internal Revenue Service, and the Economic Research Service of the Department of Agriculture and the National Agricultural Statistics Service, the Bureau of Transportation Statistics, and the Energy Information Administration. Other agencies primarily concerned with the production of statistics include the U.S. National Center for Health Statistics, the U.S. National Center for Education Statistics, the Bureau of Justice Statistics, Social Security Administration (Office of Research, Evaluation, and Statistics), and the Environmental Protection Agency. We will concentrate on the providers of economic statistics and begin with the most fundamental data, that on the population of the nation.

U.S. Bureau of the Census

The cornerstone of the U.S. Statistical system is the decennial *Census of Population and Housing*, prepared by the Bureau of the Census in the Department of Commerce. It was mandated by the Constitution and goes back to the first census in 1790. Originally designed for apportioning the House of Representatives among the states, it today serves not only this purpose but also provides fundamental information on the geographical distribution of the population, its age and sex structure, educational attainment, employment status, family income, language spoken at home, housing stock, and journey to work. Here are some questions you can answer from this fundamental source. With all of these questions, the point is not so much that you find the answer to the particular question but that you note what is available on a particular subject. To every question posed, therefore, you should add the question, “What else interesting do you see in this table?”

P1. What was the population of the United States in 1790?
All the following questions refer to 1990 or the year of latest decennial census available.
P2. What was the population of the United States?
P3. What was the population of Maryland?
P4. What was the population of Baltimore?
. P5. In their journey to work, how many Marylanders carpooled?
P6. How many Americans aged 5 and over speak Arabic at home?
P7. How many have at least a Bachelor’s degree?
P8. What percentage of occupied housing units were owner-occupied?

Alexander Hamilton, Washington’s Secretary of the Treasury, saw the need for economic information to promote the development of industry and urged the establishment of a *Census of Manufactures*. The first was conducted in 1809, five years after his death. Now also a product of the Bureau of the Census, it shows the number of establishments, employees, payroll, value of shipments, value added, and materials consumed by kind by industry and by state. Following the pattern established in the *Census of Manufactures*, the Census Bureau now conducts the *Census of Finance, Insurance, and Real Estate Industries*, the *Census of Retail Trade*, the *Census of Wholesale Trade*, the *Census of Service Industries*, the *Census of Transportation, Communications, and Utilities*, and the *Census of Agriculture*, the *Census of Construction Industries*, the *Census of Governments*. These are conducted every five years in years ending in 2 and 7. In addition, since 1949 there has been an *Annual Survey of Manufactures* to provide annual data for manufactures, but it is based on a sample and provides only some of the detail found in the *Census*.

Each of these censuses is published in many volumes of large pages and small print. Recently, they have become available on CD-ROM. The reason that the government conducts and publishes these massive censuses is quite simple: business demands them. In the early months of the Eisenhower administration, some high official got the idea that the *Census of Manufactures* was an unwanted and unnecessary burden on business that smacked of “creeping socialism.” The census scheduled for 1952, which would have been conducted in early 1953, was canceled. The result was an uproar from business and the formation of the Federal Statistical Users Group with the explicit goal to lobby for adequate statistics. The census was reinstated, but it was too late to do it for the year 1952. Instead, it was done for the year 1954. For reasons of balance in its work load, the Census Bureau wanted the economic censuses in years ending in 2 and 7. To move back towards this pattern, the next one was done in 1958. Because of the work of the decennial population census for 1960, the next was not until 1963, and finally in 1967, fifteen years after the disruption, the venerable *Census of Manufactures* got back to its intended pattern of production. It was, however, all the stronger for this attack, for it now had an organized and conscious constituency. The pressure coming from this group is probably responsible for the development of the other economic censuses.

All of these censuses define industries and products in terms of the *Standard Industrial Classification Manual (SIC)* issued by the U.S. Office of Management and Budget. The SIC uses a numerical numbering system in which two-digit industries are broad industries. For example, within Manufacturing, there are

20 Food and kindred products

21 Tobacco products

22 Textile mill products

23 Apparel and other textile products

...

39 Miscellaneous manufactures

Within a two-digit industry, there may be a number of three-digit industries. For example

20	Food and kindred products
201	Meat products
202	Dairy products
203	Preserved fruits and vegetables
...	

And within a three-digit, there may be one or more four-digit groups, for example

201	Meat products
2011	Meat packing plants
2013	Sausages and other prepared meats
2015	Poultry slaughtering and processing

Numbering codes for products extend normally to seven digits in the censuses but to only five digits in the *Annual Survey of Manufactures*. Establishments — a term that more or less means a plant at a particular location — are assigned a four-digit code corresponding to the four-digit product code which accounts for the largest share of its shipments. *Industries* are then defined as collections of establishments all having a given SIC code. Thus, we can speak of Industry 2015 Poultry slaughtering and processing, or of Industry 201 Meat products, or of Industry 20 Food and kindred products.

The SIC system is used by all statistical agencies, not just those in the Census Bureau.

Besides these censuses, the Census Bureau compiles foreign trade statistics, *The Quarterly Financial Report, The Annual Capital Expenditures Survey, Current Population Reports, Current Industrial Reports, County Business Patterns* -- the only Federal source of economic data at the county level -- and, of course, *The Statistical Abstract of the United States.*

Here are a few questions you can answer from these Census Bureau products. You can find all the answers in the *Statistical Abstract*, but look at the footnotes of the tables and report not only the numerical answer but also the original source of the information. In these questions, MRY means the “most recent year” for which data is available; in your answers, always report what this most recent year is.

C1. What is the four-digit SIC code of Paper Mills? How many establishments were there in the MRY? What were the shipments, value added, payroll, number of employees, and number of production workers of this four-digit industry in the MRY?

C2. What were retail sales in the Washington-Baltimore Consolidated Metropolitan Statistical Area (CMSA) in the MRY? How did this area compare in retail sales to the New York area?

C3. What is the SIC number of Television and radio broadcasting? How many establishments were there in this industry in the MRY? What was their revenue, payroll, and number of paid employees?

C4. What percentage of the tax revenue of state and local governments came from property taxes in the MRY?

C5. What were the stockholder’s equity and the debt of corporations in the Non-durable goods industries in the MRY?   How has the proportion of debt to equity changed since the beginning of the data?

Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) is part of the Department of Labor. As you might imagine, it is responsible for tracking employment and unemployment. It uses two sources to do so. The first is monthly reporting from establishments on the changes in their employment, hours, and earnings. The second is a household survey asking about employment, labor force participation, unemployment, and demographic characteristics such as age, sex, and ethnic background. This sample survey is the basis of the much-publicized unemployment rate.

The Labor Department is, of course, involved in labor negotiations; a key piece of information needed in these is what has happened recently to prices. Consequently, the BLS became also charged with preparing price indexes. It has two major products in this area: the Consumer Price Index (CPI) and the Producers’ Price Index (PPI). Each of these has prices on thousands of products. Combining them into a single measure led the BLS, in the case of the CPI, into conducting a Survey of Consumer Expenditures to determine appropriate weights. Originally conducted only every ten years or so, this survey has been conducted continuously since the early

1980's. The detailed results are an important source of data not only for weighting the CPI but for anyone interested in marketing to consumers or studying consumer behavior. Recently, there has been a small political storm over the measurement of the CPI because it is used to index Social Security payments and tax brackets. We shall return at the end of the course to look at some of the measurement issues involved.

Here are some questions you can answer from data collected and prepared, in most cases, by the Bureau of Labor Statistics.

BLS1. What was the average unemployment rate in Maryland in the MRY? How did that compare with the national average?

BLS2. What was the labor force participation rate of married women with children under age 6 in 1960, 1970, 1980, 1990 and the MRY?

BLS3. Which two Metropolitan Statistical Areas (MSA) had the fastest rate of growth of their consumer prices between the base period and the MRY? Which two had the slowest? What were their MRY CPI's with relative to the base period? What was the base? For the Washington MSA, which component of the index (food, housing, apparel and so on) had the fastest growth over this period? Which the slowest?

BLS4. How many economists were employed in the earliest year of data and in the MRY? What is the fastest growing occupation that you spot in the table where you find this information? Which the slowest?

BLS5. How much did the average two-person family spend (in the MRY) on clothing for women and girls? for men and boys?

BLS6. Your firm is considering opening an office in either San Francisco or Tuscon. What would be the comparative costs of housing and of living in general which middle management would face in these two cities? Note carefully the source of this data.

The Bureau of Economic Analysis

The Bureau of Economic Analysis (BEA), like the Bureau of the Census, is part of the U.S. Department of Commerce; but its mission and competence is quite different. Census is a primary collection agency; its analytical problems are mainly in assuring the quality of the data by proper survey design, proper allowance for non-response, proper checking for logical consistency in the responses, proper insurance that the privacy of respondents is protected, and similar issues. The BEA, by contrast, does very little primary data collection. Its work is primarily to create logically coherent systems of accounts from the data collected by Census and other sources. The best known of these systems of accounts are the National Income and Product Accounts (NIPA) and the Balance of Payments (BoP). It also prepares estimates of capital stock by industry, measures of personal income by state, and input-output tables showing the sales of each industry in the economy to each other industry and to various categories of final demand.

The contribution of BEA is in the comprehensiveness, coherence, comparability and timeliness of the statistics. All of these are well illustrated by the statistic for Gross Domestic Product (GDP), itself a part of the NIPA. In all the Census publications you will nowhere find a number so *comprehensive*, one that represents the whole economy so well as does the GDP.

The GDP, however, should be the same whether we calculate it from the products produced and sold to final demand as the sum of

+ Personal consumption
+ Gross private domestic investment
+ Exports
- Imports
+ Government purchases of goods and services
= Gross domestic product

or as the sum

+ Labor income
+ Capital income
+ Indirect taxes
= Gross domestic product.

Notice that the sources of data used to calculate GDP in these two ways are totally different. Conceptually, the result should be the same, but when the calculations are first done, it is hardly surprising to find a large difference between the two results. By insisting upon *coherence*, however, BEA is able to refine the data and achieve a better measure than could be obtained from one source alone.

Comparing the results of, say, the 1987 and the 1992 *Census of Manufactures* is difficult for all the statistics expressed in dollar terms because inflation has changed the meaning of a dollar's worth of output. BEA deals with this *comparability* problem by providing many series in the NIPA in both current and constant price. Finally, BEA is able to produce an estimate of the GDP and some one thousand other series that are part of the quarterly NIPA for a quarter within a month after the end of that quarter. Thus, the NIPA for the fourth quarter of 1997 appeared just before the end of January, 1998. This *timeliness* is a major factor in the importance of the GDP measure in our national culture. Of course, these early estimates are revised as more and more data become available. The final estimates for 1997 cannot be made until the 1997 economic censuses are available and BEA has had time to work over them carefully. In practice, that is likely to mean sometime in 2001. Until then, all the NIPA series since 1992 are subject to revision. In the meantime, however, BEA's current estimates provide a general indication of how the economy is doing.

BEA1. What was the GDP in the MRY? How much of this GDP went for national defense?

BEA2. What was the rate of growth of GDP in the MRY in current prices? In constant prices?

BEA3. What was the Gross State Product (GSP) of Maryland in the MRY? Maryland is, of course, a small state. How many states had a larger GSP? How many states had more GSP originating in Manufacturing?

BEA4. What were Personal income, Personal tax and nontax payments, and Personal savings in the MRY? (Nontax payments are payments to governments that are not taxes, e.g. fines or admissions to parks.)

BEA5. What were U.S. Direct investment abroad and foreign Direct investment in the United States in the MRY?

The Federal Reserve

The Federal Reserve System is the nation's central bank. Most of its functions are facilitating banking operations, regulating banks, and managing the creation of money. (The actual printing and coining of currency is a Treasury function.) Incidental to these functions, however, the Fed is the source of statistics on interest rates, foreign exchange rates, various measures of the quantity of money, the condition of banks, and consumer credit. It also prepares a set of accounts known as the Flow of Funds which is a sort of sister to the NIPA. Where the NIPA simply shows Personal saving or Business saving, the Flow of Funds accounts show what sorts of financial instruments these savings went into. Surprisingly perhaps, the Fed is also the source of the monthly Indexes of Industrial Production and Capacity. These indexes were developed because the Fed felt the need, in order to manage monetary policy effectively, to know what was happening to real output in the economy on a monthly basis.

The Fed's data enable you to answer these questions:

Fed1. The interest rate at which banks lend to one another to meet their reserve requirements at the Federal Reserve is known as the Federal Funds rate. Since most of this lending is for less than 24 hours, this rate is the most sensitive indicator of current credit conditions. What were its high and low points in the data available to you? What was its value in the MRY?

Fed2. Between 1980 and 1993, the most dynamic element component of the M1 money supply was the "Other checkable deposits". What exactly are these deposits? What has happened to them since 1993?

Fed3. When individuals save, they must put the saving into some form — a bank account or currency, a stock or bond, a new car, a new house, or something else. According to the Flow of Funds accounts, what were the principal forms chosen in the MRY?

Fed4. Between 1987 and the MRY, which two-digit manufacturing industry had the greatest growth in industrial production? How much was it? Which had the least? How much was it?

The Internal Revenue Service

Incidental to its collections of taxes, the IRS publishes a series volumes known as *Statistics of Income*, with separate volumes for individuals, partnerships, and corporations. From the individual volume you can find out:

IRS1. What was the average tax rate on reports in the $75,000 - 100,000 range in the MRY?

U.S. National Center for Health Statistics

This center publishes *Vital Statistics of the United States*. "Vital statistics" means statistics about "life" (*vita* in Latin). More specifically, they are statistics about births and deaths, but are usually extended to marriages, pregnancies, and causes of death, and disease in general. The center also publishes statistics on health care resources and health care practices. Here is just one sample of what you can learn from its publications:

CHS1. What percentage of Americans eat breakfast? What percentage are more than 20 percent overweight?

U.S. Energy Information Administration

The EIA, part of the Department of Energy, organizes data on energy production and use.

EIA1 What has happened to energy consumption per dollar of GDP since 1970?

Others

You will find in the *Statistical Abstract* yet other statistics for individual subjects. The Environmental Protection Agency has a number of statistical publications on the areas it monitors. The Department of Agriculture produces an annual volume, *Agricultural Statistics*, and many current releases related to agriculture. The U.S. National Center for Education Statistics publishes the *Digest of Education Statistics.* Virtually every federal department (except

State) has at least one statistical publication. Even the Federal Bureau of Investigation publishes an annual volume, *Crime in the United States*.

But to get the full story, ...

As you have by now realized, the *Statistical Abstract* is your guide to this wealth of information. On the other hand, it actually contains only a tiny fraction of the data which is available in the primary sources. Moreover, it is always a bit out-of-date. Many of the series are monthly or at least quarterly. To get the current data or the full detail, you must turn to the original source. These are indicated in the footnotes. Unfortunately, you will then no longer find the standardization and ease of use that the *Statistical Abstract* offers.

Using the *Statistical Abstract of the United States* in BSOS computer laboratory

The laboratory is located on the lowest flow of Lefrak. The simplest entrance is from the south, opposite the dinning hall. Enter and go straight ahead, jumping over the wall (or going around it, if you must), passing through the double doors, and continuing to the end of the hall. Turn left and go to the third room on the left, known locally as Room 3. This is where our classes will be, but the programs work from any room. Sit down at any machine. The screen should be showing two squares. Click on the one labeled “Students”. A login window appears. Click in the “Class account” check box, and give the name field as “econ321"; tap the ‘Tab’ key. Then fill in the password “stat” and tap ‘Enter’. (You can compute with the class account; later you may want to print. To do so, you must register for your own account and make a deposit. You will be charged 10 cents per page printed. After you have your own account, you will want to routinely log in using it. Your account is not charged for computing, only for printing.) Once logged in, you will get the BSOS main menu; click on “Viewers” and then on “Acrobat with Search”. The Adobe Acrobat reader will start. Click on “File” on the main menu and then on “Open”. An “Open dialog” window opens. Fill in the file name as:

```
g:\g\saus\welcome
```

and tap enter. You will follow this procedure to this point (with the change to your own account and password) every time you start work with the *Statistical Abstract*.

Now let us suppose that you want to know what you can learn about banks and banking from the *Statistical Abstract*. There are two different ways to search.

Way 1. Via the Table of Contents. Click on the square on the left labeled “Contents and Index” The table of contents appears. Find the most likely chapter, and click on the blue text of the chapter title. That chapter will open in your viewer. You can use the scroll bar at the right to scroll through the chapter. You can also click on the binoculars (without the sheet of paper behind it) to search for a particular word or phrase. The search is limited to the current chapter.

Way 2. Via indexed search. Click on the fourth button from the right on the “speed bar.” It looks like a small pair of binoculars in front of a dog-eared sheet of paper. In the window

> which appears, check the “Word-stemming” box if it is not already checked, fill in “banks” as the search target, and then click “Search”. (If the “Search” button is greyed, that is, disabled, then click on the “Indexes” button and click in the box next to the listed index, so that an x appears in this box. Then click OK. When you return, the “Search” button should be enabled.) You will get a list of all tables that contain the words “bank”, “banks”, “bankers” or “banking”. (Without the check in the Word-stemming box, you would get only “banks”.) You can use the arrow keys or the mouse to select the table you would like to look at. When the table name is highlighted, tap ‘Enter’ or double-click on the table name. When the table comes up in the viewer, all the occurrences of the search words will be highlighted. To get back to the list of tables with “hits” on the search, click on the third speed button from the right, the one that looks like a pie-chart with a dog-eared sheet behind it. On the search dialog box, there is also a Thesaurus box. Check it and search for “entertainment”; you will find that you also find appearances of “amusements” and “recreation.” On the other hand, when I looked for “pigs”, I did not find “hogs,” so the Thesaurus seems to be less than perfect. This engine also allows the use of “and” and “or” as logical operators. If you search for “cities and prices”, you will get only tables in which both words appear. If you search for “pigs or hogs”, you will get all tables in which either word appears.

Each of these ways of looking has its advantages. The first is best if you want to see what is available in a general area, because you can browse around from one table to another within a chapter easily. The second is best if you want to search the whole book. Between the two, you should be able to find answers easily to all the questions.

Please write your answers on a sheet of paper with the question numbers. At the bottom, please write and sign the statement: “Though I may have had help in learning the software, these answers represent my own work; and I could now find them without assistance.”

Chapter 2
Working with Statistics

In the first chapter, we were just finding information from the *Statistical Abstract*. You will usually want to statistics not as just an isolated number but in combination with others. For example, we looked at Gross State Product of the different states, but it might be much more meaningful to ask about Gross State Product per square mile. You will often want to show data graphically so that it can be easily grasped and remembered. In this chapter, therefore, we proceed to look at ways of presenting statistics and combining information from different tables. Our tool for doing so will be Lotus 1-2-3. The route into the data in a format in which we can work with it is to click on “Spreadsheet” link which appears in blue at the bottom of each table. Doing so takes us into 1-2-3 with a worksheet showing the data in the printed table and, in most cases much more related information. In many cases, the printed table shows data for selected years while the worksheet sheet shows it for all years available to the makers of the *Statistical Abstract*. In some cases, additional series are shown. In a few, more recent data that arrived after the book went to press are given.

Graphing data

Let us begin by drawing a graph of the recent history of some interest rates. Look for “Money Market Interest Rates”. That should bring you to a familiar table in which the Federal funds rate is the top line. Now click on the blue Spreadsheet link. Do not be alarmed if you are told that you have read access only. You would not want write access, because you might then mess up the file for others. You can save your work to the C drive of the machine you are using or to a diskette in the A drive by using the File | Save as command.

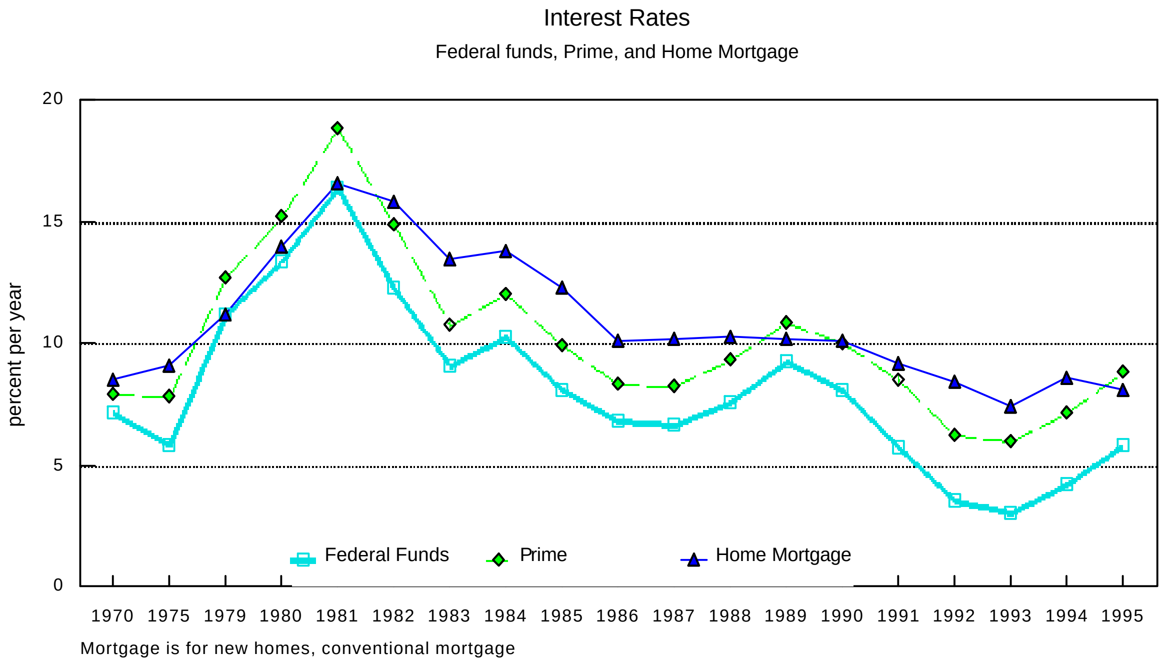
(If you do not have 1-2-3, you can substitute Quattro Pro or Excel. They have roughly the same features as 1-2-3, but I will explain the commands only for 1-2-3. To make one of these other spreadsheet programs start when you click the “Spreadsheet” link, you must associate that application with files having the .wk1 extension. First note the full path name of the application you want to start. Then start Windows Help by clicking on the “Start” button in the lower left corner of the screen, then pick “Help”, then look in the Index for “file types” and then read “To create or modify a file type.” You want to make a “file type” consisting of files with the extension .wk1, and you want your spreadsheet program to perform the “open” action when one of these files is selected. I have never actually done this, so let me know how you fare.)

We want to make a graph with three rates, the Federal funds rate, the Prime rate (the rate at which banks lend to their “best” customers), and the Mortgage rate for conventional mortgages on new homes. If you are new to spreadsheets, a few words of explanation are in order. You will notice that the screen is divided into cells, and that across the top you find the letters A, B, C, etc., while down the left side you see the numbers 1, 2, 3, etc.. A cell is designated by these two coordinates. Thus the cell in the upper left is A1, the cell to its right is B1, and cell below A1 is

A2. A *range* is a group of contiguous cells. Thus, we find the Federal funds rate in the range B12..T12, while the Prime rate is B15..T15, and the Mortgage rate is in B45..T45. (These were where they were for the 1996 book; if you are using a more recent year, the ranges will probably have extended through Column U, or V, or further, and the rows may have changed slightly.) What is the range which contains the dates?

To draw a graph, you first click Tools on the main menu and then click Chart. You get a little box that invites you to enter a data range to be charted. Enter the range for the Federal funds rate and then click OK. Your cursor has turned into something that looks a little like a graph. Put it where you would like for the graph to appear — the top left is a good place in this case — and click. A graph appears. It has, however, only one series, no dates, and a poor title. You will notice that it has little black squares in the corners. These are called "handles," and you can use them to push and pull the graph to the size you want. If you click on the spreadsheet behind the graph, the handles go away and you are just back in the spreadsheet mode. Click on the graph, and the handles re-appear. With the handles showing, you will notice that the main menu has a "Chart" item. Click on it and you will find what you need to add more data ranges, titles, and so on. The best way to learn what to do is just to play, trying this and that. A click of the left mouse button selects a part of the graph, even an individual series. A click of the right mouse button then brings up a box which allows you to set its properties. After a few minutes, I had the graph below.

Exercise 2.1. See if you can more or less recreate this graph. Save the worksheet which you create; you will need it for the exercises of the next chapter.



Exercise 2.2: Draw some other time series graph from data you have discovered in the *Statistical Abstract.* Add a comment on what you see interesting in your graph.

Bringing the graph into a document

As soon as you have your graph in 1-2-3, you will want to bring it into a word processing program. Start the word processor. (If you think you might need my help, use WordPerfect.) Then in 1-2-3 use File | Save as to save your file to a local drive, for example, your C Drive. Then click on the graph in an outer corner so that the handles appear around the whole graph. Then hold down Ctrl and tap C. This copies your graph to the Windows clipboard. Then go to the word processor, click where you want the graph, and the do Ctrl-V. In a second or so, you should see the graph appear in the document. You can still use the handles to stretch the chart if you like. If you have a great deal of data and you want to import only the graph, it may reduce the size of your document to use Edit | Paste special and paste as a "picture."

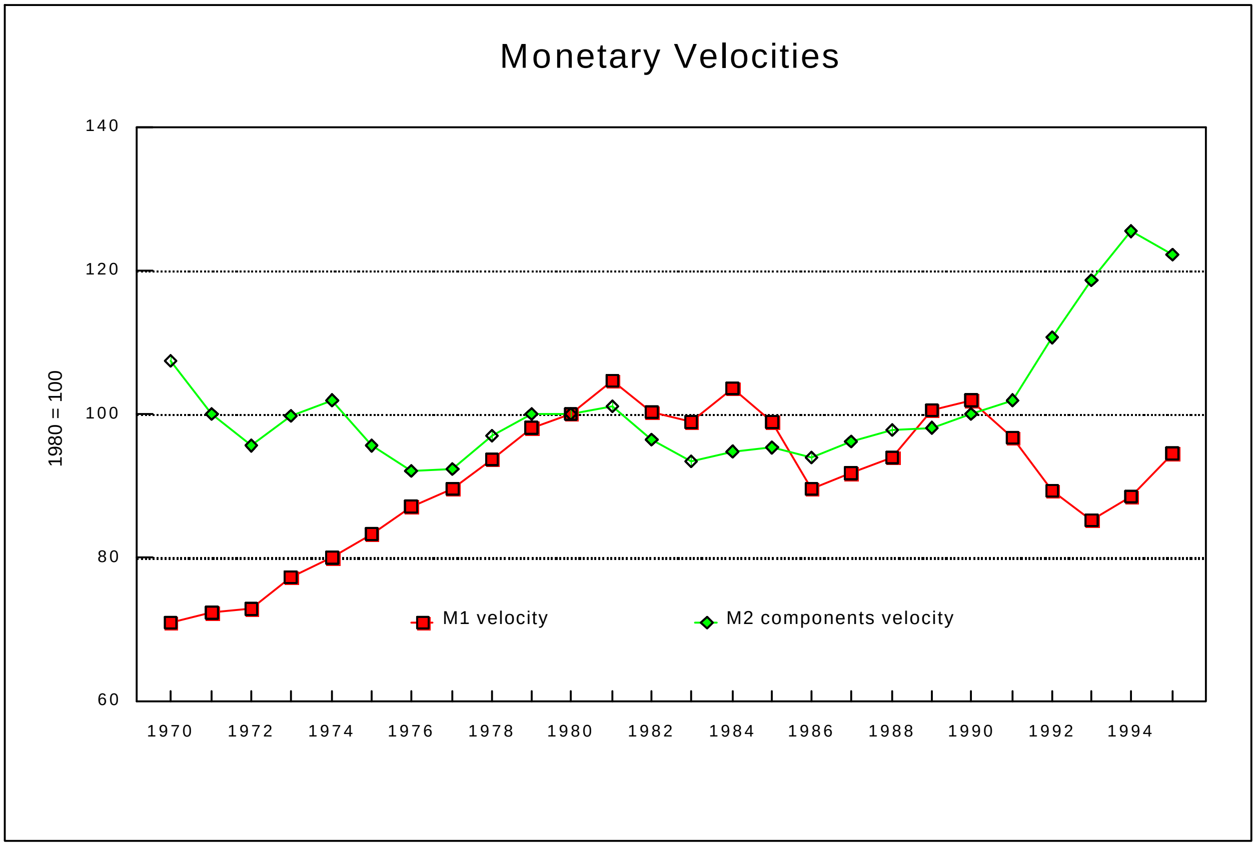
Combining data from several spreadsheets

Now let us combine data from two tables. A good example is calculation of velocities of various components of the money supply. The velocity is the ratio of GDP in current prices to particular types of money. We will use just two types, M1 and the components of M2 not in M1. (What actually is in these two components.) Find the money supply table; you can do an indexed search on "M1 and M2". Go to the spreadsheet. You will see that the data begins in 1970. Select the range that contains the dates and M1. (I find that the easiest way to "select" is with the keyboard. Put the cursor in the upper left corner of the range to be selected, hold down 'Shift' and run the cursor with the arrow keys to the right and down so that the selected cells reverse colors.) Copy to the clipboard with Ctrl C.

Now open a new worksheet by clicking File | New . Put the cursor in the upper left corner and do Ctrl-V, and your data appears in the new worksheet. (The word ITEM should appear in cell A1) Let's call this new worksheet the "computation worksheet." Click on Window in the main menu, go back to the other worksheet, select the "M2 components not in M1" and bring that data over into the computation worksheet in the line below the M1 data, namely line 5. Now click on Acrobat Reader in the bar at the bottom of the screen, and go find the series for GDP. The best series is in Table 685, where you find it back to 1960. We can use it back only to 1970, because that is where our money supply data began. Select the GDP data for 1970 to the MRY, copy it to the clipboard (Ctrl C), click on Window in the main menu, go to the computation worksheet, and copy the clipboard into the line below the two monetary components, namely, line 6.

We now have data from two tables combined into one in the computation worksheet. The next step is to compute the ratios of GDP to the two money supplies. Put the cursor in cell B8 and enter +B$6/B4 ; you should be dividing M1 by GDP.. The number 4.83022 should appear. Now copy this cell to the clipboard, select the rectangle in which it is the upper left corner and extending two lines in depth and out to the MRY to the right. Then copy the clipboard (Ctrl V) into this area. The cell references will magically change so that we get the velocity of the two types in all the years. The $ in front of the 6 in +B$6/B4 kept the 6 from changing to 7 in the lower line.

At this point you could chart the two series; but since they are of rather different magnitudes, the chart will be clearer if you first convert them to both be 100 in, say 1980. Put the converted series into two lines below the two lines of the velocities themselves. (You will need to use the device of the $ in front of the column letter of the 1980 column to prevent it from changing in the various cells.) Then graph the two lines. My result is shown below. You may have more recent data.



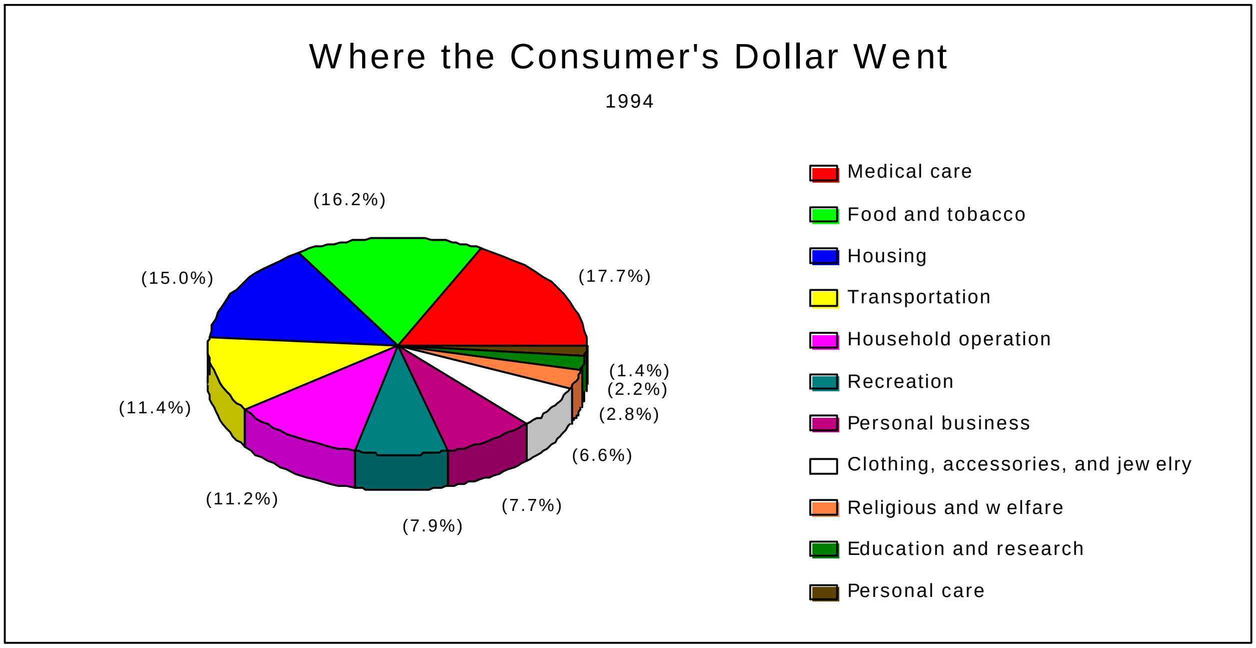
Exercise 2.3. Produce a similar graph. (Save the spreadsheet; you will need it for Chapter 3.)

This graph is, by the way, the despair of monetary economists and the Federal Reserve. Note that the M1 velocity was steadily rising up to 1981, so that it could not be used as an indicator of monetary tightness. M2, on the other hand, was giving useful indications. From 1983 to 1990, M1 is the more sensitive indicator but the general movement of the two is the same. After 1991, however, the two gave strongly conflicting indications. M2 said that money was cruelly tight; M1 said that it was almost irresponsibly easy. Which do you think was correct more nearly correct? (Look back at the interest graph.) What should the Fed be looking at in setting monetary policy? Unfortunately, there seems to be no longer an answer that can be given with confidence.

Exercise 2.4: The interest rates which you graphed in Exercise 1 are nominal and are strongly influenced by inflation. The *real interest rate* which someone making a loan at those rates would have received is the nominal rate minus the rate of inflation. Graph the nominal and real Federal Funds rate as well as the rate of inflation over the same period as in Exercise 1. (That is three series in one graph.) For the inflation rate, use the percentage change in the Consumer Price Index for urban households (CPI-U). (Look for Consumer Price Indexes. You will find a table with the rates of change already computed, but the years run down the page instead of across. To compute the real rates, you will need to *transpose* the range containing the information. The necessary 1-2-3 is Range | Transpose; but before trying to do it, you should perhaps read 1-2-3 Help on the subject. Search its index for “Transpose”.) Were the years of high nominal rates also the years of the high real rates? How would you describe the effect of inflation on interest rates? (Save the spreadsheet; you will need it for the next chapter.)

Exercise 2.5. Produce a different table or graph with data from two different tables. If you think of nothing better, you can compute Gross State Product per square mile. You can experiment with making a map that shows your data. Comment on what you find.

Needless to say, you can also make pie charts. Here is one showing shares of the consumer’s budget in 1994. The table in the *Statistical Abstract* had far too many data series for a pie chart. So I copied it to a new worksheet, cut out the columns except the names and 1994, then selected the detailed sections that I did not want and tapped the ‘Delete’ key. That left a table with many blank lines. I then selected the whole area with data, the did Range | Sort, and did a descending order sort on the column with data. Then I selected the are with data (now much smaller), Clicked on Tools | Chart, then changed the type to 3D-Pie, and filled in a title. Here is the result.



Exercise 2.6: Make a pie chart with data of your own choosing. Be sure to comment about what you see.

Chapter 3

Statistics from the Internet

Many statistics are now available from the Internet. These sources are more up-to-date than the *Statistical Abstract* and often more complete. They lack, however, the good indexing and uniformity of presentation that makes the *Abstract* such a good starting place. Once you are acquainted with the broad outlines of available data, however, you can brave the chaotic presentation of Federal statistics on the Internet. Fortunately, there is a web site that provides some degree of indexing of federal data. It, however, only provides links to sites maintained by individual agencies. Among them, there is no uniformity. One might suppose that federal agencies all answerable in some way to the Office of Management and Budget could agree on the format for presenting, say, a monthly statistical series. But such is by no means the case, as you will notice. For our purposes, however, that will not be a major problem. Moreover, the Internet sites are continually changing. What you find may not be exactly what I found, so if the instructions here prove erroneous, use your own ingenuity to find and use the data.

Our task in this chapter is simply to update the interest rate and velocity graphs from the previous chapter. We will need to get the data subsequent to that we already have for:

Consumer price index, percent change

Federal funds rate
Prime rate
Mortgage rate

GDP

M1
M2 (from which you compute M2-M1)

These are listed in approximate order of difficulty of getting from the net.

After starting as usual, from the BSOS main menu select "Info and Internet Access". Then start "Netscape Communicator 4.02". We begin from the "central" federal site. In the "Location" of Netscape, put

http://www.fedstats.gov

and then tap 'Enter'. This is the one Internet address you should memorize for access to federal statistics. Knowing it will save you lots of time — and spare you lots of advertisements — on the Internet search engines. (The corresponding address for international data is dsbb.imf.org . Have a quick look at it.) When Fedstats starts, pick "A to Z" which opens an alphabetical list of topics. Find Consumer Price Indexes and click on the link. Note where you are at this point, that is, what Internet address is showing in the Location box. Does it make sense in terms of what you know about the producer of the Consumer Price Indexes? The first item on the list offered you, the CPI Summary, has a table of percent changes over each calendar year, just what you want. Note them down, or if you have the worksheet with your graph open, enter them directly into the worksheet.

Now to get the Interest Rate data. Use the "Back" button to back up to A to Z on Fedstats, find Interest rates, and follow the link. You will be taken to www.bog.frb.fed.us. (Bog is presumably for Board of Governors, and frb is — somewhat redundantly, for Federal Reserve Board) The releases listed on this page are for just very recent data. Follow the link, near the top of the page, to "Historical Data". There you find that you are on another index page.   Locate the series you want and follow the link to "annual data." There you will find the interest rates to update your graph. Move them into your worksheet.

Now for GDP. Click on Back until you get back to the A to Z page again. Go after GDP. The first link will take you to a press release with only quarterly data. You need annual data for 1996 and 1997. Note the message:

> Links to other pages on
> our web site are shown at
> the bottom of this page.

Follow that link and select "NIPA data". On the page where that puts you, you will need the link opposite the "More comprehensive data" heading. (You can use either the HTML or the PDF format; the HTML works faster.) When you land there, have a good look around. You are looking at the official version of the National Accounts of the United States. All other presentations are derivative. You should have no difficulty finding GDP for the years you need.

To update the money supply variables, use the Back button to get back to "A to Z", find Money stock, follow the link, and then on Fed's page take the link to Historical data. You will find M1 and M2 monthly, but, alas, not the annual data you need. I do not think it is on the Intenet. You need to get the data into a spread sheet so that you can sum up and average the monthly data for the last two years. Select the data you want — you may as well select it all so that you get the headings — and copy it to the clipboard with Ctrl C. Copying it directly into 1-2-3 or Excel does not work; all the data goes into the first column. Instead, you have to copy it as a plain text file to your computer and then import it into the spreadsheet. In Netscape, on "File | Save as" of the main menu, save the file, say, to your diskette in the A drive. Start 1-2-3 if you have not already done so and get a clean worksheet; you can do so by "File | New". Then do "File | Open", select the file you have just copied to your machine, and tap enter. On the screen that then opens, select "Automatically parse based on file layout". The file will then be correctly imported. Use the @sum() function to sum up the 12 months of the years you need -- and divide by 12 -- to get the annual data you need. Copy the results to the clipboard and move them over to the worksheet with the graphs.

You are now in a position to extend the ranges of the graphs for another two years. Looking at the velocities, has money been tight or easy? What has happened to real interest rates? Does the story these two graphs tell fit with what you would have expected from economic theory? Why or why not?

Chapter 4

Introduction to the Public Use Microdata Sample

Turning to the source

So far in this course, we have been looking at statistics which had been much processed before we got them. We now want to turn to looking at data closer to the source, closer to the primary collection. We cannot actually gather any significant volume of data ourselves, but we can do the next best thing. We can look at that data as gathered by the U.S. Census Bureau from individual households and made available as a Public Use Microdata Sample (PUMS) on CD ROM.

In connection with the decennial census of population, the Census also gives to a sample of households a "long form" questionnaire which asks about the age, sex, education, language, work, journey to work, earnings, and income of each person in the household and about a number of characteristics of the house, including its estimated value. As originally collected, these data contain the name of the person and address of the house. In the PUMS data, all such identifying information has been removed to protect the respondent. Only the state, a broad area within the state, and a few characteristics of the neighborhood — rural, suburb, central city — remain. From the returns of the long forms, the Census makes up what is intended to be a 1% sample. The quality of the data has proven sufficiently good that Census also feels able to release a 5% sample. The 1% sample, however, provides vastly more data than we need to illustrate the elementary concepts in this course, and we will stick with it. Since one of our concerns is to study what can be said on the basis of a *sample* of data, we will, in fact, draw 20 subsamples from the Census 1% sample, each of them being 1/1000 of the 1% sample and having roughly 2,500 persons. You may very well be the only student working with your sample, so your work is important to the class effort to see how much answers differ because of differences in the sample.

To keep the worksheets you will be using to manageable proportions, I have selected about thirty items of information for each person in our subsamples and have written these samples as text files with the names of sample0.dat, ..., sample19.dat. You will find them in the G:\G\PUMS directory. You can copy your sample to a diskette and take it home. It is a text file that any spreadsheet program should be able to read. Instructions here, however, are for 1-2-3. Once you get it into the program, you will see short titles of columns across the top. These are derived from the full titles shown in the appendix of this chapter, which also appear as the file codes.txt in the above-mentioned directory.

One of the main problems in using primary data is what to do about missing data, the questions the respondent did not answer. Census has to some extent solved this problem for us in PUMS. If a question was unanswered, an answer has been "allocated" to it on the basis of answers to the same question in the responses of similar families. In the PUMS data as it reaches us on the CD ROM from the Census, these allocations are all flagged so that the user can decide what to do about them. For our purposes, however, we can certainly accept the Census allocations, so I have omitted the allocation flag fields.

As you can imagine, it is easier to get responses to long forms from some types of families than from others.   Well-educated, English-speaking families are more likely to respond than poorly educated families with limited knowledge of English. Thus, when Census does get a response from a family with low probability of response, it should be given a higher weight than a response from a family with a high probability of response. The weights which Census believes are appropriate for each person and each household are indicated in PUMS. I have included these weights in the samples, but using them would complicate our computations. Since we are aiming to illustrate concepts, we will ignore the complications these weights would add. If you were interested in obtaining the most representative results possible, you should use them.

A look ahead

When you get your sample and bring it into the spread sheet program, look it over. What can you see? The answer is apt to be something like, "Thousands of details and nothing in general." That is a good beginning, for much of statistics is devoted to how to say something meaningful, memorable, and useful on the basis of a mass of information which, in its initial state, numbs the mind by its quantity. In the next chapter, the fifth, we will study ways to describe a *single* variable such as income. We will see ways to describe the variable by its distribution function, its frequency function, or its Lorentz curve. We will go on to look at measures of central tendency such as the mean, mode, and median of the variable. We will also develop a measure — called the standard deviation — of the dispersion of the variable. We will look at these measures not only for specific samples but also for certain mathematical distributions, including the normal curve that plays a remarkable role in statistics.

These measures, as you can readily imagine, will turn out to be different for different samples. That fact difference raises the important question What can we say about the whole population on the basis of a sample? Making valid statements about the population on the basis of a sample is called *statistical inference* and will occupy us in Chapter 6.

In chapter 7, we will look at ways to find what relation exists between one variable and a number of explanatory variables working together. The method used is called, for historical reasons, *regression* analysis. We will explore a number of “tricks of the trade” in regression, including dummy variables, proxy variables, and logarithmic transformation of variables.

The measures which express the relations found by regression are different for each sample, so the population values are subjects for statistical inference. In chapter 8, we will look at how that may be done, but we must stress that the results here must be much more qualified than in the case of the mean.

In the last section of the course, we will return in chapter 9 to look at some of the systems of statistics, in particular at the National Accounts, the Consumer Price Index, and money supply statistics. Each of these has important conceptual matters that must be explored. In my view, working out of these concepts is every bit as much a part of economic statistics as is the regression and sampling theory developed in the earlier chapters.

Finally, in chapter 10, we apply regression analysis to the time-series statistics studied in chapter 9 to look for macroeconomic relations. In particular, we will explore the determinants of investment in equipment, and you will have the adventure of finding a role for interest rates among those determinants. Equations such as you estimate in this chapter can be combined to produce a comprehensive macroeconomic model of the economy. That step, however, is beyond the scope of this book; it is treated in detail in my *Craft of Economic Modeling.*

Thus, you will have experienced by the end of this course the four major aspects of the use of statistics to study the economy:

Finding data
Description and analysis of microdata, often called cross-section analysis
Principles of construction macroeconomic data
Analysis of the relations to be found in the time-series of macroeconomic data.

Chapter 5

Statistics of a Single Variable

Distribution and density functions

Bring your sample into your spreadsheet. In 1-2-3, you do File | Open, give the name of you sample file (e.g. sample0.dat) and tell the program that the file is ASCII text. You should get a worksheet with names, one per column, across the top. I had some trouble getting these names to come in correctly; please be sure that there is one per column. (You can cut and paste if necessary.)

Now take a good look at the data. What do you see? Probably all you can say is something like, "I can't see the forest for the trees," or "Too much data to see anything!" Good. Much of statistics is about how to find something meaningful in such a mass of data. In this Chapter, we will work on the description of a single variable; in later chapters, we turn to relations among variables.

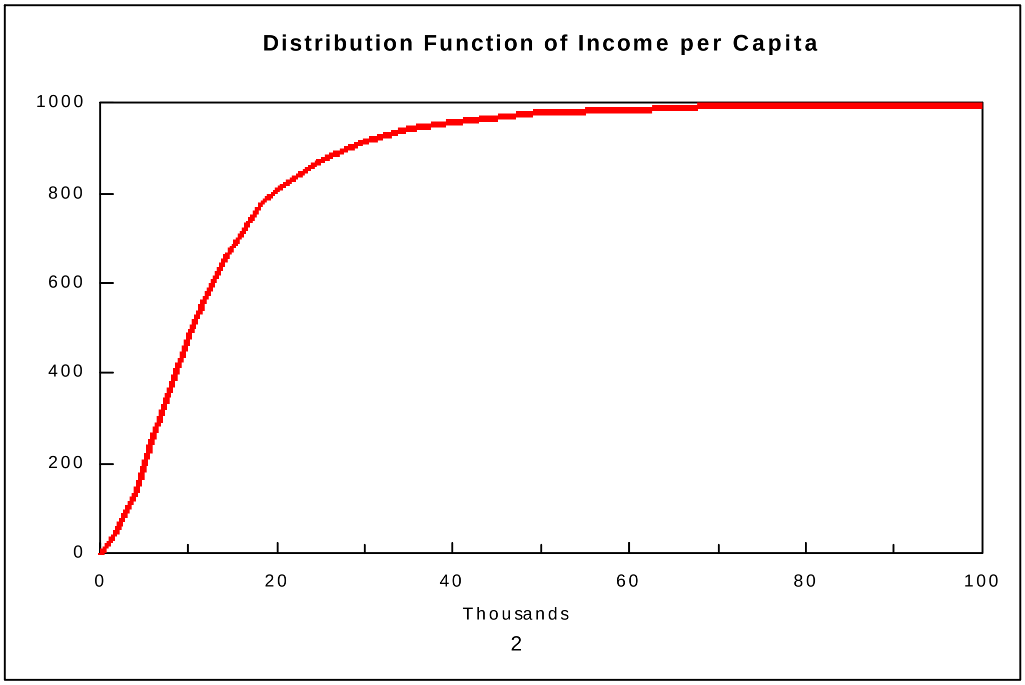
The variable that I want to work with is income per capita within households. For example, a four-person household with a combined income of $40,000 would be said to have a per capita income within the household of $10,000. This income would seem a better measure of the welfare of a person than either total family income (without regard to how many mouths there are to feed or feet to clad) or than the income of the person (which may be zero for children or the spouse providing child care in the home). This measure does not appear as such in the PUMS data, but at the far right of the spreadsheet is a column for household income and a few columns to the left appears the number of people in the household. Divide one of these columns by the other. Check over the column of results to be sure it contains no error messages, which appear as ERR instead of the numeric result. My first sample had a few such errors due, I believe, to a problem in making the PUMS CD. I simply removed the offending observations from the sample. You may also notice a number of 0's in the results. Apparently, no household income was reported for institutionalized individuals or others living in group housing -- such as a fraternity house. Where such individuals had income and were reported as living in households of size 1, I put down the individual's income as the household income.

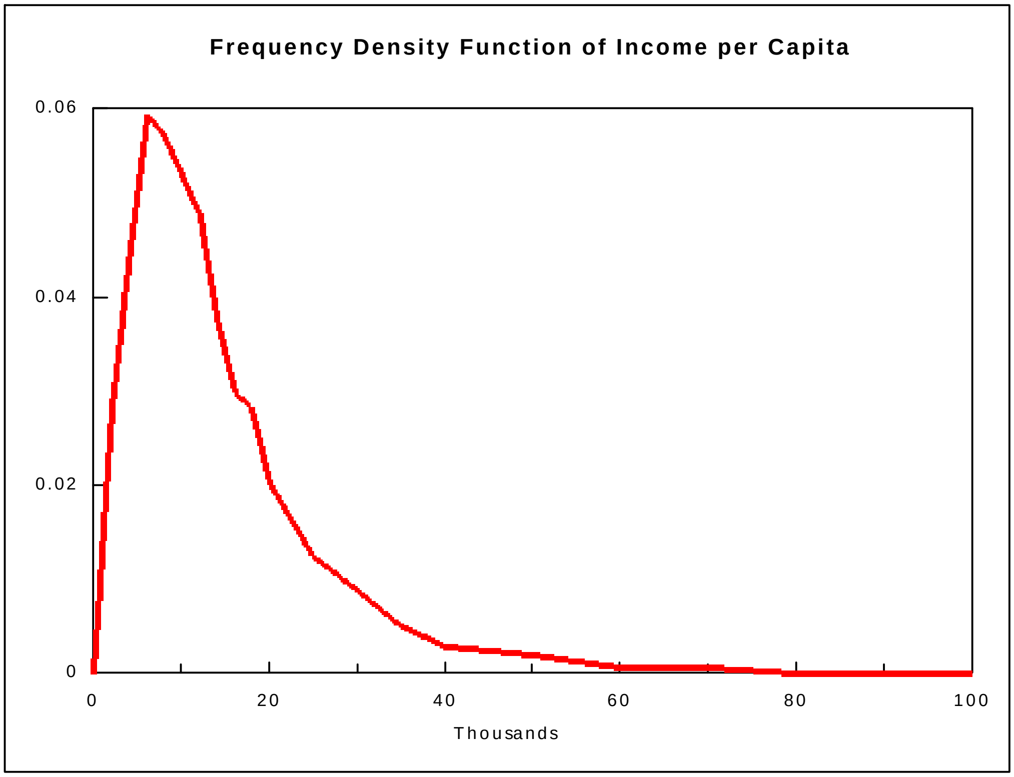
The next step is to move the column of data over to another worksheet where you can work on it without danger of messing up the basic data. Select the range of cells with the data on per capita income and copy it to the clipboard with Ctrl C. (To select the range, you may find it useful to know that you can put the cell cursor on the top entry in the column, hold down Shift, tap End and then ↓, and the cursor will move to the cell before the first blank cell, which will be in effect the end of the range to be selected.) Then use File | New to open a blank worksheet. Put the cell cursor in cell A3 (to leave room for a title in the top two lines) and do Edit | Special paste. On the form which then appears, check "Copy formulas as values," and click OK.

Now let us find out how income was distributed over the population. Type the first column (the one labeled “Bracket”) of the table below into Column F of your worksheet. (I have plans for Columns B-E.) This column shows brackets for making up the income distribution. Now choose Range | Analyze | Distribution from the main menu. A form appears asking for the data range and the bin range. The bin range is the range where the brackets appear; the data range is the range in column A where the data is. Fill them in and click OK. In the column to the right of the bin range, the counts of the number of individuals in each bracket appear. How similar is your result to mine shown in the table below?

Bracket	Count	Permil	Distribution	Delta X	Density
-3000	0		Function		Function
-1	4				
0	23	0.0	0.0		
2000	135	56.9	56.9	2000	0.02843
4000	206	86.8	143.6	2000	0.04339
6000	282	118.8	262.4	2000	0.05939
8000	273	115.0	377.4	2000	0.05750
10000	252	106.1	483.6	2000	0.05307
12000	231	97.3	580.9	2000	0.04865
14000	180	75.8	656.7	2000	0.03791
16000	142	59.8	716.5	2000	0.02991
18000	134	56.4	773.0	2000	0.02822
20000	96	40.4	813.4	2000	0.02022
25000	148	62.3	875.7	5000	0.01247
30000	104	43.8	919.5	5000	0.00876
35000	61	25.7	945.2	5000	0.00514
40000	36	15.2	960.4	5000	0.00303
50000	48	20.2	980.6	10000	0.00202
60000	19	8.0	988.6	10000	0.00080
70000	16	6.7	995.4	10000	0.00067
80000	3	1.3	996.6	10000	0.00013
90000	3	1.3	997.9	10000	0.00013
100000	1	0.4	998.3	10000	0.00004
125000	3	1.3	999.6	25000	0.00005
200000	1	0.4	1000.0	75000	0.00001
	0				

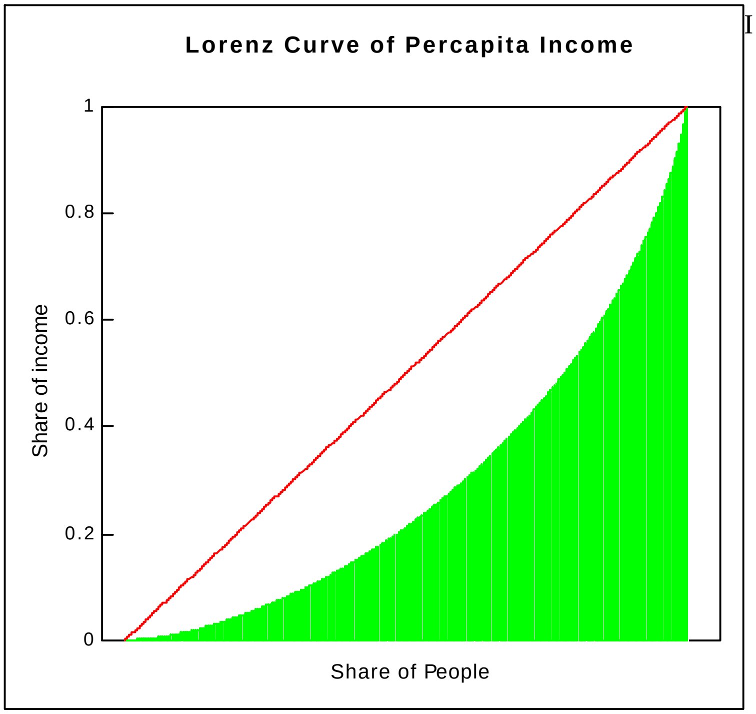
One reads the counts as the number of people with income above the next lower bracket entry but less than or equal to this one. Thus, the table shows 135 people with income above 0 but less than or equal 2000. The absolute number of people in each cell depends, of course, on the sample size. It is easier to get comparable number by dividing each cell of the Counts column by the total number of persons and multiplying by 1000. This result appears in the Permil column. (Permil is like percent, but is per thousand instead of per hundred.) The next column, labeled Distribution Function is the cumulation of the Permil column. That is, it starts at zero and with each bracket adds the Permil entry for that bracket to the previous entry in the Distribution Function column. The distribution function is shown in the first graph below.





There is still a problem in comparing the cells of the permil column, and therefore in graphing it, because some cells are broader than others, that is, they cover a broader range of income. The column labeled Delta X shows the range of income in each bracket, and the column labeled Density function shows the results of dividing the Permil column by the Delta X column. Just as we speak of the number of people per square mile as the density of the population, so this density is the permil of the population per dollar of the range of incomes, or, to put it in other words, it is the fraction of the population per $1000 of the income range. Its graph is shown in the second figure above. These two functions, distribution and density, show essentially the same information in two different ways. They play a central role in the theory of statistics, as we shall soon see.

Before turning to that theory, however, we should show a third way of presenting the information which is particularly revealing when applied to income. This is the Lorenz curve shown below. For any x between 0 and 1, it shows the fraction of total income received by people in the lowest x*100 percent of the population. For x = .25, for example, it shows the fraction of income received by the quarter of the population with lowest income. If everyone had the same income, the Lorenz curve would be the diagonal line. One commonly used measure of income inequality is the Gini coefficient, which is the ratio of the area between the diagonal and the Lorenz curve to the whole area under the diagonal. It is zero for a perfectly equal distribution and 1 for the most unequal distribution possible.



To make the Lorenz curve, you arrange the incomes in column A of your worksheet in increasing order. (Select the range, do Range | Sort, give A4 as the item to be sorted on, and click OK to do the sort.) Next, cumulate column A in column B. Then put into column C column B divided by its last cell. Into Column D, put the integers 0, 1, 2, 3, etc. (Put 0 in the first cell, then put in next the formula to add 1 to the cell above; then copy that formula into the entire range.) Then in column E, divide column D by its bottom element. Finally, make an XY chart with column C as the Y range and column E as the X range. Draw the diagonal by adding a second Y range as column E.

Exercise 5.1. Construct the distribution and density functions for per capita income within families for your sample. Draw them.

Exercise 5.2. Construct the Lorenz curve for your sample.

So far we have just used the distribution function and the frequency density function to describe the data in our sample. It is, however, but a short step to imagine that there are such functions for the whole population. In doing so, we move from the concrete to the concept, from numbers to theory. We shall call the theoretical distribution function F(x) and the theoretical density function f(x). There is, however, a slight difference. Our distribution function rose from 0 to 1000, as is convenient for showing in a table. By convention, F(x) rises from 0 to 1.000, so it is our function divided by 1000. For plotting our function, we did not need to bother about mathematical properties like whether the function was continuous and differentiable. For F(x) and f(x), however, we shall assume that they are both continuous and that at least F(x) is differentiable. Indeed, we defined f(x) by

$$f(x) = (F(x + \Delta x) - F(x))/\Delta x$$

and if we now take the limit as Δx goes to 0, we see that f(x) is simply the derivative of F(x), or, in symbols, $f(x) = F'(x)$. We shall do more with these functions in the next section.

Measures of central tendency: means, medians, and modes

While the graph of a distribution or frequency function conveys much more information than can be summarized in a single number, it is natural to want to have some single number to indicate a typical value, some sort of "middle" of the distribution. There are three candidates commonly used: the mean, the median, and the mode.

The mean is the arithmetical average. For the sample of incomes, one just sums up all the incomes and divides by the number of people. For the sample used here, the sum of the incomes was $32,260,468 for 2374 people with non-zero income, giving a mean income of $13,589.

The median is the income of the person in the middle of the Lorenz curve, the one with as many people above as below. With our sample, it came out at an even $8,800.

The mode is the income for which the density function is the highest, the "fashionable" income, so to speak. When working with actual data, we only have the density function for intervals. From the table above or the graph we can see that modal interval is $4,000 to $6,000. The density for the $6,000 to $8,000 interval, howerer, is almost as high while that of the $2,000 to $4,000 interval is well below, so it would appear that the mode is close to $6,000.

These rather large differences in the three measures are the result of the strongly asymmetric shape of the density function. Note, for example, that doubling the income of everyone in the upper half of the distribution would have no effect on the mode and the median, but would have a a big effect on the mean.

Not surprisingly, the mean is the most commonly used of these *measures of central tendency*, but each of them has its use.

The mean can be computed not only from the raw data but also from the distribution or frequency functions. If we take many points, x_i, along the x axis, bracket points if you will, and use the definition

$$\Delta x_i = x_{i+1} - x_i$$

then the mean, m, is approximately

$$m = \sum_i \frac{F(x_i + \Delta x_i) - F(x_i)}{\Delta x_i} x_i$$

so as we let the number of points go to infinity and the size of the intervals go to zero, we get

$$\mu = \int_{-\infty}^{\infty} f(x)\, x\, dx \,.$$

Example: The uniform density function. The simplest density function is f(x) = 1 for 0 <= x <= 1 and f(x) = 0 for all other x. The mean of the uniform distribution is

$$\int_0^1 x dx = .5 x^2 \Big|_0^1 = .5 * 1 - .5 * 0 = .5 \;.$$

Exercise 5.3. (a) Determine the constant *c* which makes the following function a density function, that is, such that its integral from minus infinity to plus infinity is 1.0:

$$f(x) = c(x^3 - 6x^2 + 9x) \quad for\ 0 \le x \le 3;\ f(x) = 0\ for\ other\ x.$$

(b) What is the mean of this distribution? What is the mode?

A measure of dispersion: the standard deviation

The measures of the central tendency tell us something about where the middle of the distribution is, but they tell us nothing about how spread out it is. We could invent several measures of dispersion, but one is so much more used than all the others that we will rest content with it. It is known as the standard deviation, is usually denoted by σ, and is defined as

$$\sigma = \sqrt{\sum_{i=1}^{N} (x_i - \mu)^2 / N} \;.$$

In words, the standard deviation is the square root of the average of the squared deviations from the mean. The square of the standard deviation is called the *variance*. In terms of the theoretical frequency function, it is

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \;.$$

Exercise 5.4. Compute the standard deviation of income per capita within families for your sample. Do it first by using no @function from your spreadsheet program except @sum() and @sqrt(); that is, take the deviations from the mean, square them, sum them, and so on. Next, look up "statistical @functions" in the help files of your program. In 1-2-3, you will find that @std() will take the standard deviation of a population, while @stds() estimates the standard deviation of the population from a sample. Use the former for the moment.

Exercise 5.5. Compute the variance and standard deviation of the uniform distribution.

Exercise 5.6. Compute the variance and standard deviation of the distribution studied in Exercise 5.3.

The fraction of the population (or sample) for which $|x - \mu| > k\sigma$, which we shall call $P(|x - \mu| > k\sigma$)is always less than $1/k^2$. This remarkable result, known as the Chebyshev (or Tschebichieff) inequality, holds for any distribution. It says, for example, that less than one quarter of the population can be more than 2 standard deviations from the mean and less than one ninth can be more than 3 σ from μ. Thus, the standard deviation is a quite generally valid measure of how spread out the distribution is.

The Chebyshev inequality is easy to demonstrate. Let R be the range of values of x for which $|x - \mu| > k\sigma$ and R' all the other values. Then

$$\sigma^2 = \int_R (x - \mu)^2 f(x)dx + \int_{R'} (x - \mu)^2 f(x)dx \geq k^2\sigma^2 \int_R f(x)dx .$$

The first equality is just the definition of σ^2; the second follows from the facts that (a) in the first integral $(x - \mu)^2 > k^2 \sigma^2$ for x in R, while (b) the second integral is certainly not negative. The integral on the far right is just $P(|x - \mu| > k\sigma$), so dividing both sides by $k^2\sigma^2$ gives $P(|x - \mu| > k\sigma) \leq 1/k^2$, which is the Chebyshev inequality. You will notice in the proof that the inequality is quite likely satisfied with lots of room to spare.

Exercise 5.7. What fraction of the people in your sample had per capita incomes within one standard deviation of the mean? Within two? Within three?

The normal distribution

When one begins to cast about looking for simple mathematical functions that will be density functions over a wide range of x values, not a lot functions rush to mind. The uniform density function is clearly a pretty special case that will not describe the distribution of many variables. The polynomials, such as used in Exercise 3, have a limited use as density functions because it is tricky to keep them from turning negative in the desired range; if the 6 in the function of Exercise 3 is changed to 7, the function will turn negative inside the range of x from 0 to 3. The function c/x for some constant c and x >= 1 is not a candidate because its integral is infinity and a density function must have an integral of 1. The function c/x^2 has a finite integral for the same range of x so c can be chosen to make it a density function, but the mean is infinite. Similarly, the function c/x^3 can be a density function and has a finite mean for this range of x but the standard deviation is infinite. These functions are useful as brain-teasers but are of limited value for describing economic phenomena. The function f(x) = c e$^{-|x|}$ can be a density function and has finite mean and variance, but its shape, with a kink at x = 0, seems a bit odd.

It is with a sense of relief, therefore, that one comes upon the function

$$`(x) \;=\; ce^{-x^{2}/2}$$

The reason for the ½ is to give the function a variance of 1, as we shall verify. From the symmetry of the function around x = 0, we see that the mean is zero. Moreover, with the constant c defined so that $\int_{-\infty}^{\infty} f(x)dx \;=\; 1,$ we can use integration by parts to calculate the variance:

$$\sigma^2 \;=\; c\int_{-\infty}^{\infty} x^2 e^{-x^{2}/2}dx \;=\; -cxe^{-x^{2}/2}\Big|_{-\infty}^{\infty} \;+\; c\int_{-\infty}^{\infty} e^{-x^{2}/2}dx \;=\; (-0 \;+\; 0) \;+1 \;=\; 1.$$

Here we used the usual integration by parts formula

$$\int udv \;=\; uv \;-\; \int v d\iota$$

with

$$u \;=\; x \quad and \;\; dv \;=\; xe^{-x^{2}/2}dx.$$

You can use l'Hopital's rule to show that $\quad x\;e^{-x^{2}/2} \;\rightarrow\; 0 \;\; as \;\; x \;\rightarrow\; \infty.$

Exercise 5.8. Use your spreadsheet program to compute and graph the normal curve. In column A, put the numbers -5.0, -4.9, ..., +4.9, +5.0. (Put -5 in A1, +A1-.1 in A2, and copy A2 to A3..A102.) In column B put the values of exp(-x^2/2) for the values of x in column A. Sum them up and divide by 10, the range of the values of x. You should get something just a shade less than 2.506628, which is $\sqrt{2\pi}$ to six decimal places, as you may readily verify with the @pi function in your spreadsheet. You total is a shade less because there is a tiny area under the normal curve for values of x larger than 5. In column C, put the values of column B divided by their total. Graph this column C.

You may certainly take the result in Exercise 5.8 as strong evidence that the proper value of *c* in the formula for the normal curve is $1/\sqrt{2\,\pi}$. If, however, you are surprised to find π, which you thought had to do with circles, jumping out to meet you in this context, you may be interested in the mathematical evaluation of the integral -- and if you are not interested, you may skip this evaluation. It uses an unusual trick, namely, instead of evaluating the integral, we evaluate the square of half of it. Let the upper half of the integral we wish to evaluate be

$$I = \int_0^\infty e^{-x^2/2} dx.$$

Then

$$I^2 = \int_0^\infty e^{-x^2/2} dx \int_0^\infty e^{-y^2/2} dy = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)/2} dx dy$$

To evaluate the multiple integral on the right, we convert from the rectangular coordinates (x,y) to the polar coordinates (r, θ) where r is the length of the line from the origin to the point (x,y) and θ is the angle (measured in radians) which this line makes with the x axis. The original x and y are related to r and θ by the equations

x = r sinθ
y = r cosθ

so that in particular $x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$. In polar coordinates, the area element which is dxdy in rectangular coordinates becomes $rd\theta dr$, because the length of the arc swept out by an increment in θ of $d\theta$ is $rd\theta$. Thus, transformed to polar coordinates the above multiple integral becomes

$$I^2 = \int_0^{\pi/2} \int_0^\infty e^{-r^2/2} r dr d\theta = \int_0^{\pi/2} -e^{-r^2/2} \Big|_0^\infty d\theta = \int_0^{\pi/2} d\theta = \pi/2,$$

where the second equality follows from the integration by parts we did in computing the variance. We want *c* such that *2cI = 1* or

$$c = 1/2I = 1/2\sqrt{\pi/2} = 1/\sqrt{2\pi}\ .$$

The normal curve which we have studied so far has μ = 0 and σ = 1. We can easily generalize the normal to have any specified μ and σ:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This frequency function is much used. A variable with this density function is said to be N(μ,σ).

Exercise 5.9. Verify that a variable that is N(μσ) in fact has a mean of μ and a standard deviation of σ.

Chapter 6

Sampling

Random samples and random variables

A great deal of statistics involves making inferences about a population on the basis of a sample. A lot can be learned about, say, the within-family per capita income of Americans, on the basis of a properly selected sample of as few as fifty persons. But the words "properly selected" must be emphasized. If the sample is taken in a shelter for homeless people or from among the attendees at a $1000-per-plate political fund-raising dinner, the results are unlikely to tell us much about the total American population. The validity of the sample as an indicator of the population depends upon its selection by a *mechanism that has no known or suspected connection with the values of the variables to be observed.* The problem with the sample drawn in the homeless shelter is that people are there precisely because their income is low, while attendance at the fund-raiser is also connected to income.

A sample that has been selected in a way unconnected with the values to be observed is called a *random* sample. The value of a variable, say income, observed on a selected individual is then called a *random* variable. Notice carefully that in calling the income of successive persons in the sample a random variable we by no means mean to imply that chance has much if anything to do with the person's income. Indeed, we shall later try to find its connection with age, occupation, education, sex, race, and so on. Neither need the sample selection process involve chance; the census taker may give the long form in a perfectly mechanical way to every tenth person she counts. The order in which she comes to them is determined by where they live. And few people, if asked why they live where they do would answer, "Just by chance." The only requirement is that the mechanism would not tend to exclude or include people in the sample because of the values of the variables being collected.

In PUMS, as we have noted, Census believes that some households were more likely to have been included than others and has provided us with weights to correct for this bias. Most of them are close to 1 for the households. We will continue to ignore these weights and *consider PUMS a true random sample*. I do not mean to recommend this neglect of the weights as good statistical practice, but as just a simplification that allows us to concentrate on basic concepts at this early stage of your experience with statistics.

Independent variables

The *joint density function* of two variables, *x* and *y*, is the function *f(x,y)* such that $f(x_0,y_0)\Delta x\Delta y$ is the fraction of the population with x lying between x_0 and $x_0 + \Delta x$ and y lying between y_0 and $y_0 + \Delta y$. Mathematically, two random variables are said to be *independent* if f(x,y) can be written g(x)h(y) where g(x) and h(y) are the density functions of x and y. This mathematical definition is simple enough, but what does it correspond to in reality? When can we suppose that two random variables are independent? The answer must be something like the definition of "random." If we cannot imagine how knowing the value of x would help us in forecasting y, given that we already know the density function, h(y), from which y comes, then we may say that x and y are independent. Successive tosses of a coin are usually thought to be independent. Values of successive houses down a street, given only the distribution of house values in the whole city, are probably not independent. We are going to assume that our sampling procedure has given us samples in which the values of a given variable, say the per capita income, are independent random variables. We then have two important theorems; in both of them x and y are independent random variables, *x* has density function *g(x)* with mean μ_x and variance σ_x^2 while *y* has density function *h(y)* with mean μ_y and variance σ_y^2 .

The mean of a sum of independent random variables, x and y, is the sum of their means.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x + y)f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x + y)g(x)h(y)dxdy = \int_{-\infty}^{\infty}xg(x)dx + \int_{-\infty}^{\infty}yh(y)dy = \mu_x + \mu_y.$$

The variance of a sum of independent random variables, x and y, is the sum of their variances.

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x + y - (\mu_x + \mu_y))^2 f(x,y)dydx &= \int_{-\infty}^{\infty}(x-\mu_x)^2 g(x)dx\int_{-\infty}^{\infty}h(y)dy \\ &\quad + 2\int_{-\infty}^{\infty}(x-\mu_x)g(x)dx\int_{-\infty}^{\infty}(y-\mu_y)h(y)dy \\ &\quad + \int_{-\infty}^{\infty}g(x)dx\int_{-\infty}^{\infty}(y-\mu_y)^2 h(y)dy \\ &= \sigma_x^2 + 2(\mu_x-\mu_x)(\mu_y-\mu_y) + \sigma_y^2 \\ &= \sigma_x^2 + \sigma_x^2.\end{aligned}$$

Though we wrote out the formulas for sum of two variables, the reasoning clearly applies just as well to n variables. If we now suppose that all the n variables have the same density function with mean μ and variance σ^2, as they would if they are all determined by random drawings from the same population, then the mean of the sum is nμ and its variance is $n\sigma^2$. Now the variance of *ax*, where *a* is a constant and *x* is a random variable with variance σ^2, is easily seen from the definition of variance to be $a^2 \sigma^2$. Thus, if we divide the sum of the *n* independent random variables by *n* (so that the *a* of the previous sentence is *1/n*), we get a variable with mean μ and variance $n\sigma^2/n^2 = \sigma^2/n$. Thus, we have the extemely important result that the mean of a sample of n values on a random variable is the same as the mean of the population from which it is drawn but *the variance of the sample mean is only one nth the variance of the variance of the population*. That factor of

1/n is the whole reason why large samples are better than small samples for determining the mean of the population.

Let us illustrate this important relation with the data from the distribution of per capita income. The mean from my sample of 2400 persons with positive income was $13,589. The standard deviation of this sample was $12,525. Thus, the standard deviation of the mean of a sample of 1 would be $12,525; for a sample of 2, the standard deviation of the mean falls to 8856; for a sample of 9 it is $4195; for a sample of 50, it is $1771; for the sample of 2400 it is 256, and for the for the whole 1 percent sample it is 8. Increasing the sample size is a good example of decreasing returns to scale. Increasing the size from 1 to 2 reduces the standard deviation by nearly 30 percent. Two more observations are needed to get it down another 30 percent; then four more to get it down another 30 percent, then eight more, and so on. On the other hand, the cost of collecting the sample, aside from some fixed planning costs, increases more or less linearly with the size of the sample. When a sample survey is being planned, this tradeoff between the more or less constant marginal cost of enlarging the sample and the declining marginal benefit of each added observation must be carefully weighed.

An experiment in sampling

Suppose that we drew a random sample of 50 people from a population with an income distribution such as that found in the previous chapter. If we know nothing about the population except what we learn from our sample, our best guess of the mean of income of the population would certainly be the mean income of our sample. But this sample mean would be a random variable; it would depend on the sample. How accurate would it be as an indicator of the population's mean? In other words, what sort of distribution would it have, and in particular what would be its standard deviation? (Remember: we don't know the standard deviation of the population.)

Since we in fact have a sample of some 2500 from PUMS, let us try to answer these questions by drawing a number of random samples of 50 from it, computing the mean income in each, and then looking at the distribution of these means. (The point of this experiment is not to determine the mean of the population but to see how the mean of small samples is distributed.)

As before, copy the column of income per capita within households to column A of a new worksheet. The function @rand produces a pseudo random number with a uniform distribution. Put @rand in cell B1 and copy it to the all the rows of column B where there are entries in column A. You will notice that when @rand is copied, the copies have different values. Since sorting also causes the values of @rand to change, it will reduce confusion if we copy column B to column C using Edit | Paste special | Copy formulas as values. Now use Range | Sort to sort this whole range on column C. The first 50 incomes are now a random sample of income, as are the next 50, and so on.

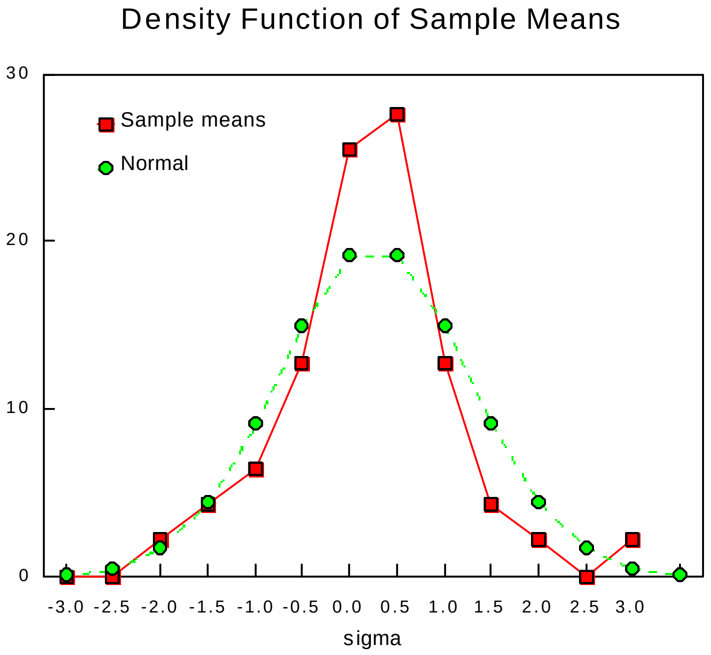
Now compute in columns D and E the means and standard deviation in each sample. To do so, put @avg(A1..A50) into D50 and @std(A1..A50) into E50. Copy these two cells into D100, D150, etc.. Now copy Columns D and E to F and G using Edit | Paste special | Copy formulas as values to do the copy. Now sort the range composed of columns F and G in descending order on column F. That will put the means and standard deviations to the top of the worksheet where they are easier to see than they are when spread out over 2500 lines.

Finally, we are ready to find their distribution. To do so, you will recall, we need a ‟bin” range. We will make it in Column I with the help of the ‟sigma units” which you should enter in Column H copying the first column of the following table. Compute the mean and the standard deviation of your whole 2500-observation sample. Divide this standard deviation by the square root of 50 to get the standard deviation of the mean of samples of size 50. Call this ‟sigmamean”. Now make up the ‟bin” range in Column I as the mean plus ‟sigmamean” times the entry in the same row of the ‟sigma units” column. Your results should look generally like the Brackets column in the following table, though the numbers will be somewhat different because you have a different sample. With the ‟bin range” ready, do Range | Analyze | Distribution to get the count in each cell. These counts will automatically go in Column J. Convert it into percent in Column K. Graph the results. The result for my sample is shown by the solid line marked with squares in the next graph. To add the normal to your graph, you can make use of the @normal(x) function, which gives the value at x of the distribution function of the Normal(0,1). For x, use the values from the ‟In sigma units” column.

In sigma	Bracket	Count	Percent	Normal	Normal
units				density	0.00000
-3.0	9375	0	0.0	0.1	0.00135
-2.5	10089	0	0.0	0.5	0.00621
-2.0	10802	1	2.1	1.7	0.02275
-1.5	11515	2	4.3	4.4	0.06681
-1.0	12228	3	6.4	9.2	0.15866
-0.5	12942	6	12.8	15.0	0.30854
0.0	13655	12	25.5	19.1	0.50000
0.5	14368	13	27.7	19.1	0.69146
1.0	15081	6	12.8	15.0	0.84134
1.5	15794	2	4.3	9.2	0.93319
2.0	16508	1	2.1	4.4	0.97725
2.5	17221	0	0.0	1.7	0.99379
3.0	17934	1	2.1	0.5	0.99865
		1		0.1	1.00000

The central limit theorem

Something rather strange, beautiful, and at first perhaps mysterious seems to be at work in the table and graph just reported. The density function of incomes was very far from the normal, but the density function of the means looks surprisingly like the normal. For comparison, the normal frequency function is shown in the graph by a dotted line with points marked by circles. Let me remind you that we introduced the normal function just because it was about the simplest function we could imagine that would make a nice density function. But here it emerges that it is a fairly good approximation of the distribution of the sample means from a decidedly non-normal population when the sample size is 50.



We see here an illustration of the working of the *central limit theorem*, arguably the most famous and most useful result of mathematical statistics. According to it, the distribution of the sample mean from a wide range of distributions approaches the normal distribution as the size of the sample increases. This remarkable result was stated by Laplace in 1809 and first given a rigorous proof by Liapounov in 1912. You may find the details in most any mathematical statistics book such as Paul G. Hoel's *Introduction to Mathematical Statistics*. Reading the proofs requires a knowledge of the series expansions of logarithms and patience to follow through the algebra. We will not go into them here. Even if we did, we would still have to ask How quickly does the distribution of the sample mean approach the normal. And for that question, we are thrown back to experiments such as we have just done.

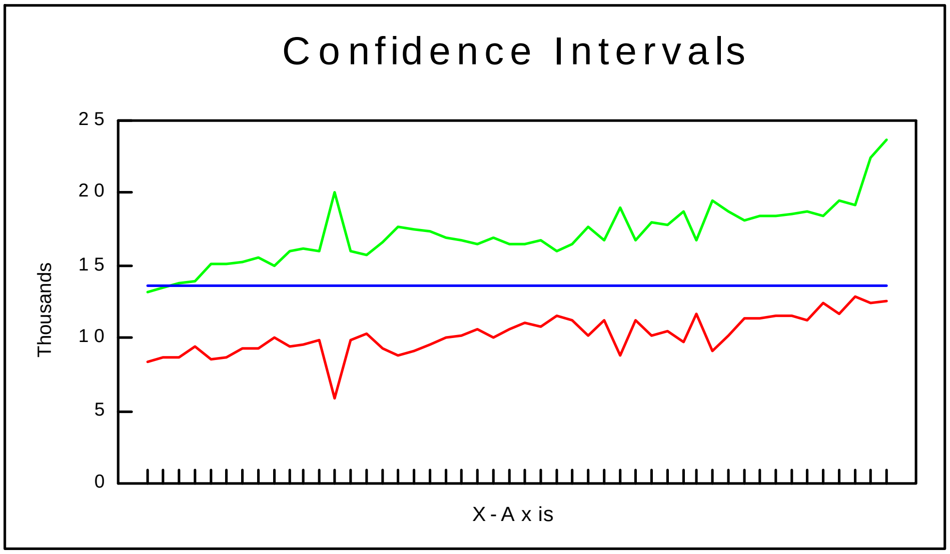
Confidence intervals

Suppose that we are given one and only one of our samples of size fifty. What does it enable us to say about where the mean of the parent distribution is? Its mean is an estimate of the population mean, but it is almost certainly not the population mean. Can we calculate an interval likely to include the population mean? More precisely, what is the narrowest interval which will include the population mean in a specified fraction of all possible samples? That fraction is called the *confidence* level for the interval and the interval is called a confidence interval.

If we can assume that the sample is large enough that the sample mean is essentially normally distributed, we can answer the question. For the specified confidence level, C, find the value of t such that C percent of the normal distribution is within $t\sigma$ of the mean. Then construct the interval as $m - ts$ and $m + ts$, where m is the sample's mean and s its standard deviation. If we are sufficiently close to the normal, then this interval should cover the population mean in about the fraction C of the cases.

From the column labeled Normal of the table above, we can see that t = 1 corresponds to a C of about .68 and t = 2 corresponds to a C of .9545. For C of exactly .95, t should be 1.96.

The graph below shows the .95 confidence intervals for the 50-observations samples. The horizontal line is the "population" mean. The interval covered the population mean in 46 of the 48 samples, or in .958 of the cases, satisfactorily close, I should think, to the theoretical expectation.



Exercise 6.1. Perform the calculations illustrated here for your PUMS samples. Do you find any striking similarities or differences from the results with my sample?

Chapter 7

Least Squares Regression

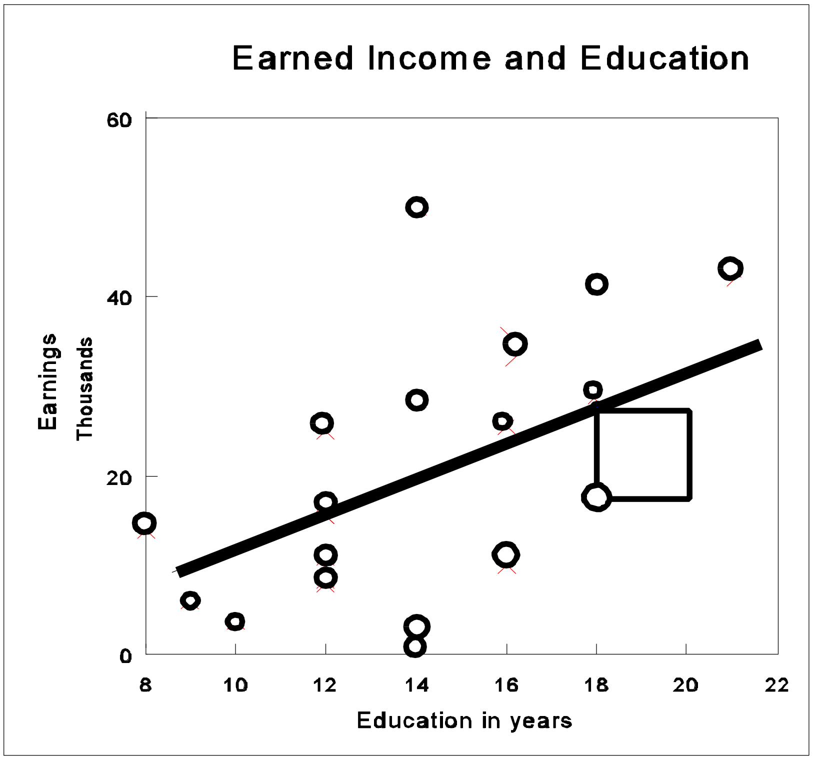
In this chapter, we will be concerned with how to find relations among variables. We shall ask, for example, What is the effect of education and age on earned income? First, we concentrate on the mechanics of this curve-fitting process. We will come back to the statistical properties of the results, and to the important question of what makes a sensible equation.

1. What is the method of least squares and why use it?

In our PUMS data, we have the earned income of each individual and a number of characteristics of that individual. Can we use these other characteristics to explain income?

Let us start with one factor. A number of you are may be in this course because you are under the impression that education influences income. Is that assumption borne out by the data you have? How can we find the relation between the two?

The simplest way to answer that question is to plot income against education shown below for twenty earners. You can then draw a line more or less through the middle of the plot, as I have tried to do. (My line is just drawn by eye; if you think another would fit better, please draw it.)



Mathematically, what drawing the line does is to determine b_1 and b_2 of the following equation:

(1) $y_t = b_1 + b_2 Ed_t$

where y_t = Earned income of person t
Ed_t = Education of person t
b_1, b_2 = constants to be determined so that the line fits the points.

This method of "eyeing in" the line has important advantages over other methods. It allows us to see what we are doing. It makes it easy to spot outliers in the data and to avoid putting a heavy weight on them. It is, however, seldom used. Theorists don't like it because no elegant theorems can be proved about it. Practitioners shun it because it is easy to see that it is subjective; more sophisticated methods allow one to cloak subjectivity in a more opaque garment.

The real problem with the method, however, arises when we realize that income depends on more than just just education. Suppose that we want to take into account the age of the earner, then we should have to estimate

(2) $y = b_1 + b_2 Ed + b_3 Age$

(The b_1 and b_2 of (2) may, of course, be different from those of (1).) To estimate these b's by eye in this equation, we need a three-dimensional construction to suspend the points in space. Then perhaps we might use a plane of light to find the plane which fits best. But what was simple for one variable becomes something of an engineering feat for two. But now, of course, we realize that earnings may also depend on the sex of the earner, thus:

(3) $y = b_1 + b_2 Ed + b_3 Age + b_4 Sex$

Now we need a four-dimensional structure of some sort; and when you have constructed it, be prepared to tackle a five-dimensional device, for earned income may also depend on disability, so that we should be estimating:

(4) $y = b_1 + b_2 Ed + b_3 Age + b_4 Sex + b_5 Disability$

Here we seem to be pretty much past hope of making a physical picture to aid us in fitting the equation to the data. If we ever want to estimate an equation with that many parameters -- and for all practical purposes, with any more than two parameters -- we must find some way to replace vision by calculation. That switch can be dangerous, and we shall frequently find that we must check to be sure that the blind calculations have not lost their way.

To find a way to calculate the parameters, let us return to the simplest case, equation (1) and the above figure, and let us agree to choose b_1 and b_2 so as to minimize the sum of the squares of the errors. The figure shows one of these squares. We twist and push the line through these points until the sum of the areas of the squares is the smallest. You may ask "Why the squares? Why not just get the smallest sum of absolute values of error?" Strangely, it is easy to compute the b's to minimize the sum of squares, but quite difficult to minimize the sum of absolute errors, and it is really this ease that is responsible for our preference for the squares.

2. How to calculate the least squares regression coefficients

In this section, we leave the specific example of the preceding section to develop a mathematically general way finding the b's that minimize the sum of the squared errors. Later, however, we will return to precisely the example described in section 1. In the previous section, we used mnemonic names like Ed, Age, and Sex for the independent, right-hand-side variables. To simplify our notation, we shall from now on in this chapter refer to them as x_1, x_2, x_3, etc; the dependent variable will continue to be y, but with no intention to mean income. To find the "least squares" values of the b's with just education as an explanatory variable, we just minimize with respect to b_1 and b_2 the following summation:

$$(5) \qquad S = \sum_{t=1}^{T} (y_t - (b_1 x_{t1} + b_2 x_{t2}))^2$$

where T is the number of observations which we have, $x_{t1} = 1$ for all t and x_{t2} = education of person t. The sigma, Σ, of course, indicates summation, and the t=1 below it indicates that the summation begins with that value of t and extends to t = T, shown above the sigma. Now if S is to be minimal with respect to both b_1 and b_2, then it must be minimal with respect to b_1 with b_2 held constant.   First, the derivative of S with respect to b_1 (with b_2 held constant) must be zero:

$$(6) \qquad \frac{\partial S}{\partial b_1} = \sum_{t=1}^{T} 2(y_t - (b_1 x_{t1} + b_2 x_{t2}))(-x_{t1}) = 0$$

Likewise, S must be minimal with respect to b_2, so

$$(7) \qquad \frac{\partial S}{\partial b_2} = \sum_{t=1}^{T} 2(y_t - (b_1 x_{t1} + b_2 x_{t2}))(-x_{t2}) = 0$$

We may see the implications of (6) and (7) more clearly if we will divide through by 2, move the terms not involving b_1 and b_2 to the right hand side, and factor out b_1 and b_2 from the sums involving them. Upon so doing, (6) and (7) become

$$
(8) \qquad \begin{aligned} b_1 \Sigma\, x_{t1}x_{t1} + b_2 \Sigma\, x_{t1}x_{t2} &= \Sigma\, x_{t1}\, y_t \\ b_1 \Sigma\, x_{t2}x_{t1} + b_2 \Sigma\, x_{t2}x_{t2} &= \Sigma\, x_{t2}\, y_t \end{aligned}
$$

(Here we have dropped the limits on the summation simply because they are always the same.) Now bear clearly in mind that the x's and y's are known, observed values. Hence, all the sums are known numbers. The unknowns are b_1 and b_2. So we see that we have two linear equations in two unknowns. You solved equations like that in high school; if perhaps you have forgotten how, I shall remind you in a moment. Given that we know how to solve linear equations, we can consider our problem of how to find the b's to be solved.

So far, you may say, we have only dealt with the two-variable case, handled satisfactorily graphically. What about more variables? Well, we might as well go for the general case and consider n independent variables x_1, ..., x_n, and the equation

$$y_t = b_1x_{t1} + ... + b_nx_{tn}.$$

Then let

$$S = \sum_{t=1}^{T} (y_t - (b_1x_{t1} + \ldots + b_nx_{tn}))^2$$

Differentiating with respect to b_1, b_2,...,b_n gives

$$
(9) \qquad \begin{aligned} \frac{\partial S}{\partial b_1} &= \sum_{t=1}^{T} 2(y_t - (b_1x_{t1} + ..\ .+ b_nx_{tn}))(-x_{t1}) = 0 \\ \ldots &= \ldots \\ \frac{\partial S}{\partial b_n} &= \sum_{t=1}^{T} 2(y_t - (b_1x_{t1} + \ldots + b_nx_{tn}))(-x_{tn}) = 0. \end{aligned}
$$

These equations may be rewritten as

$$
(10) \qquad \begin{aligned} & b_1 \Sigma\, x_{t1}x_{t1} + b_2 \Sigma\, x_{t1}x_{t2} +...+ b_n \Sigma\, x_{t1}x_{tn} = \Sigma\, x_{t1}y_t \\ & b_1 \Sigma\, x_{t2}x_{t1} + b_2 \Sigma\, x_{t2}x_{t2} +...+ b_n \Sigma\, x_{t2}x_{tn} = \Sigma\, x_{t2}y_t \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & b_1 \Sigma\, x_{tn}x_{t1} + b_2 \Sigma\, x_{tn}x_{t2} +...+ b_n \Sigma\, x_{tn}x_{tn} = \Sigma\, x_{tn}y_t \end{aligned}
$$

Do you see the system? In the first equation, the first factor in every product is x_{t1}, in the second equation, x_{t2}, and so on. In the ith column, the second factor in each sum is x_{ti}. These equations are called the *normal* equations of regression. (This use of the word "normal" has nothing to do with the "normal" distribution, but is related to calling a line perpendicular to the tangent to a curve at a point a "normal.")

Let us take a simple example with T = 5.

	t	x_{t1}	x_{t2}	x_{t3}	y_t
	1	1	10	5	17
	2	1	5	1	10
(11)	3	1	0	6	12
	4	1	10	3	16
	5	1	0	10	20

You should now verify that the equations (10) are

$$\begin{aligned} 5b_1 + 25b_2 + 25b_3 &= 75 \\ 25b_1 + 225b_2 + 85b_3 &= 380 \\ 25b_1 + 85b_2 + 171b_3 &= 415 \end{aligned}$$

Table 7.1 shows how to solve them systematically. The first three lines (L1, L2, L3) show the original equations. The next three show the results of eliminating b_1 from the second and third equations. To obtain them, first get the coefficient of b_1 in the first equation to be 1.0 by dividing the equation by b_1's coefficient, namely 5. The divisor is called the pivot element and is underlined in the table. The result is L4. Now to eliminate b_1 from equation 2, multiply the equation in L4 by b_1's coefficient in L2 and subtract from L2 to get L5. Since equals multiplied by the same thing are equal and equals subtracted from equals are equal, L5 also represents an equation. L6 is similarly calculated and is also an equation. Any and all b's which are a solution of L1-L3 are also solutions of L4-L5. We have now completed one *pivot* operation and eliminated b_1 from all equations except the first. In the next three lines, L7, L8, and L9, we similarly get 0 for the coefficient on b_2 in all but the second equation. Finally, in L10 - L12, we get zero coefficients on b_3 in all but the third equation. The last three lines are thus three equations whose solution is exactly the same as the solution of the first three line; but the solution of the last three is obvious:

$$b_1 = 4.95 \qquad b_2 = .61 \qquad b_3 = 1.40.$$

Line	b_1	b_2	b_3	=	1	Derivation
L1	5.	25.	25.		75.	
L2	25.	225.	85.		380.	Original Equations
L3	25.	85.	171.		415.	
L4#	1.	5.	5.		15.	L1/5
L5	0.	100.	-40.		5.	L2 - 25*L4
L6	0.	-40.	46.		40.	L3 - 25*L4
L7	1.	0.	7.		14.75	L4 - 5*L8
L8#	0.	1.	-0.4		.05	L5/100
L9	0.	0.	30.		42.	L6 -(-40)*L8
L10	1.	0.	0.		4.95	L7 -7*L12
L11	0.	1.	0.		.61	L8-(-.4)*L12
L12#	0.	0.	1.		1.40	L9/30

A # after a line number marks the line computed first in each panel of three lines.

Table 7.1: Least Squares Computations

The process we have followed to fit an equation to the given data known as *ordinary least squares* or *linear regression* and the b's we have found are the *regression coefficients*. The particular method of solution of the equations is known as Gauss-Jordan reduction.

Historical note on regression by least squares

The idea of fitting an equation by minimizing the sum of squared misses first appears in Adrien Marie Legendre's *Nouvelles méthodes pour la détermination des orbites des comètes* in 1805. Legendre had been involved, beginning in 1795, in the measurement of the meridian arc from Barcelona to Dunkirk, the measurement on which the length of the meter was based. In the 1805 book on comets, he seems to have discovered the method near the end of the writing and in an appendix returned to a problem in measuring the meridian arc to illustrated the new method. The method spread rapidly in astronomy and geodesy. In 1809, Carl Friedrich Gauss published a small volume in Latin on the orbits of planets; in it he not only used the method of least squares but claimed to have been using it since 1795. There is no particular reason not to believe this claim since Gauss did so much other original work in mathematics. On the other hand, he might have forgotten to tell anyone else about it had Legendre not done so first. Gauss also connected the method to the normal distribution by noting that, if the errors were normally distributed, least squares would give the estimates that maximized the probability of observing the sample which was actually observed. This connection stimulated LaPlace in his studies leading to the central limit theorem.

The term ‟regression” came from a study of ‟hereditary stature” by Francis Galton in England in the 1880's. After collecting data on the heights of about 800 relatives, he found, first, that the average of the men's heights, 68.25 inches, was 1.08 time the average of the women's heights. In all the following operations, he multiplied the women's heights by 1.08. With this adjustment, he found that if he plotted the average height of parents on the horizontal axis and the height of their offspring on the vertical axis, the “regression” line cut the 45-degree line at the average height but had a slope of 2/3 — not 1.0 — so that the heights of the children seemed to ‟regress ‟ towards the mean, or ‟mediocrity” in Galton's word. He as found that if he plotted the ‟stature” of one sibling on the horizontal axis and that of other siblings on the vertical axis, the line again had a slope of 2/3. If he used the height of only one parent, the regression was more marked, with a slope of only 1/3. With nieces and nephews on the vertical, the ‟regression ratio” was 2/9 and for grandchildren, it was 1/9.

Galton also studied the distribution of the residuals from his ‟regresssion” lines and found that ‟every one of the many series with which I have dealt in my inquiry conforms with satisfactory closeness to the ‛law or error.'” This ‟law of error” is what we would call the normal distribution. Of it, he said in a presidential address to the Anthropological Institute:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ‟law of error.” A savage, if he could understand it, would worship it as a god. It reigns with serenity in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. (‟Hereditary Stature,” *Nature,* January 28, 1886.)

He is talking about the central limit theorem, and he himself is no doubt the savage he had in mind.

Neither Galton nor his mathematical adviser at Cambridge thought of using the method of least squares to fit the ‟regression” lines. Instead, he used a method like that used by surveyors in making contour maps and observed that the contours of equal density were ellipses. His “regression” lines were the principal axis of the ellipse. His work, however, attracted wide attention and the term “regression line” became firmly established. The step forever linking the words ‟regression” and ‟least squares,” was taken by George Udny Yule, about 1897. For a full account of all these developments, see Steven M. Stigler *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge, 1986.

Exercise 7.1 Calculate the regression equation for the following data.

x1	x2	x3	y
1	5	2	14
1	4	3	13
1	6	3	17
1	7	5	20
1	6	7	19
1	8	6	21

Exercise 7.2. Check you work by using the regression command of your work sheet. See the "Regression" topic in the help files of the spreadsheet you are using for detailed instructions. With 1-2-3 release which I am using, the X-columns must all be together, side-by-side, while the Y-column can be elsewhere. The command is then Range | Analyze | Regression and you are asked to fill in the range for the X and Y variables, specify where the output (the regression coefficients and some other things) should go, and indicate whether or not the Y-intercept should be computed or set to zero. I generally put the output below the data where there is no danger that it will overwrite anything. On the question of the Y-intercept, if you have included a column of 1's in the range, as shown in the example, then you should select "Set to zero." Alternatively, you may leave out the column of 1's but select "Compute Y-intercept". The resulting coefficients are the same either way.

3. Some measures of how well the equation fits

Now that we have computed the regression coefficients, we may well ask How well does the equation fit the data? To answer, we first need to compute the values of the dependent variable "predicted" by the equation. These predicted values are denoted by $\hat{y}$ thus

$$\hat{y}_t = \sum_{i=1}^{n} b_i x_{ti} .$$

They are shown in the third column of Table 7.2 below, where the actual values are shown in the second column. The misses, or "residuals",

$$r_t = \hat{y}_t - y_t$$

are shown in the fourth column, labeled *r*. Note that the sum of the residuals is zero; it will always be zero if there is a constant term in the equation. Since we were trying to minimize the sum of squares of these residuals, this quantity is naturally of interest:

$$S = \sum_{t=1}^{T} r_t^2 .$$

Actually, the Standard Error of Estimate

$$SEE = \sqrt{S/T}$$

is easier to interpret for it has the same units as the dependent variable. Indeed, we could describe it as sort of average error.

Another measure of closeness of fit is the ratio of S, the sum of squared residuals, to D, the sum of the squared deviations from the mean of the dependent variable, ỹ. These deviations are shown in the column labeled *d* in Table 7.2. This D is, to be explicit,

$$D = \sum_{t=1}^{T} (y_t - \tilde{y})^2 .$$

t	y	$\hat{y}$	r	r^2	d	d^2	f	f^2	%
1	17.	18.05	1.05	1.1025	2.0	4.	-1.65	2.7225	6.18
2	10.	9.40	-0.60	0.3600	-5.0	25.	1.95	3.8025	6.00
3	12.	13.35	1.35	1.8225	-3.0	9.	-2.10	4.4100	11.25
4	16.	15.25	-0.75	0.5625	1.0	1.	-0.30	0.0900	4.69
5	20.	18.95	-1.05	1.1025	5.0	25.	0.00	0.0000	5.25
			-----	------	---	---		------	----
			0.00	4.9500	0.0	64.		11.0250	33.37

SEE = $\sqrt{4.95/5}$ = .995 MAPE = 33.37/5 = 6.67

R^2 = 1 - (4.95/64) = .9227 RBARSQ = 1 -(4.95/2)/(64/4) = .8453

DW = 11.025/4.95 = 2.2272 RHO = (2 - 2.2272)/2 = -.1136

Table 7.2: The Fit of the Equation

The ratio S/D would be zero for a perfect fit and 1.0 in the worst possible case (provided there is a constant term in the equation.) Since it seems a bit strange to have zero as the perfect score, the S/D ratio is subtracted from 1.0 to form what is called the "coefficient of multiple determination" or "R square" or RSQ for short,

$$R^2 = 1 - S/D$$

A "corrected" R^2, written with a bar over the R and pronounced "R bar square" is often used to allow for the fact that as more and more independent variables are added to a regression, S goes to zero. In fact, if there are as many independent variables as there are observations, the fit will normally be perfect. For the corrected R^2, therefore, S is divided by T - n, the number of observations less the number of independent variables. The formula is

$$\bar{R}^2 = 1 - \frac{S/(T-n)}{D/(T-1)},$$

where n is the number of independent variables, counting the constant term. Where it is not convenient to use superscripts, we call this measure RBARSQ. The number T - n is called the number of "degrees of freedom" in the regression.

An alternative measure of closeness of fit is the mean absolute percentage error, or MAPE, defined by

$$\text{MAPE} = 100 * \sum_{t=1}^{T} | r_t / y_t | / T$$

Its calculation is illustrated in the last column, labeled *%,* of Table 7.2.

Although RSQ, RBARSQ, and MAPE are all dimensionless pure numbers, there is no absolute standard of "how good is good" for them. Generally, any equation whose dependent variable has a strong trend will have a high RSQ while equations with volatile, untrended dependent variables will
have low RSQ. These measures are really useful only for comparing the fit of one equation for a particular variable with another equation for the same variable.

Columns 8 and 9, labeled *f* and f^2, and the calculation of *DW* are included for future reference. Do not trouble yourself with them at the moment.

Exercise 7.3: For the data in exercise 2.1, calculate also the SEE, RSQ, RBARSQ, and MAPE. Compare with the measures reported in your spreadsheet program.

4. Regression of earned income on education and other variables

Let us now return to our samples from PUMS and do a real regression. One of the economically interesting variables in the data is certainly earned income, REarning. Let us see how well we can explain it with other variables in the bank, such as education, sex, age, hours worked, and general part of the country in which the individual lives.

To begin with, bring your sample into a new spreadsheet, so as not to mess up your previous work with PUMS. Since many individuals are not in the labor force at all, we first remove from the sample individuals who have no earnings. Select the whole of the sample except the variable names at the top and then sort in descending order on earnings. Drop down in the sample to where those with zero earnings begin to appear. That will be about half way down. Select all columns of all these individuals with zero earnings and tap the 'Del' key. You are left with only those individuals that had earned income. They should be about half of the original sample. It doesn't matter for the regression, but it will convenient later to have the sample in a random order, so put in column of random numbers, copy it to the clipboard, and then do Edit | Special paste | Copy formulas as values to copy the random numbers as values over themselves. Sort the entire sample by the random numbers. Finally, to make room for the regressions, insert 26 columns to the left of the data.

Copy the desired dependent variable column, the one with the label REarn, into column D; we will need columns A, B, and C later. Put a column of 1's into column E. (Put a 1 into E2, +E2+1 into

E3, copy E3 to the clipboard, and paste the clipboard to the whole column as far down as the data goes.

Decoding a variable

Our next job is to get the years of schooling into column F. We encounter a problem here because the variable labeled YearsSch is not really years of education, but codes indicating years of education as shown in the table below. To convert it to something approximating years of schooling, we must use a table look-up function. I used @VLOOKUP(x,range,offset) in 1-2-3. Here, *x* is the code we are looking up, *range* is the location of the table, and *offset* is the column to be used.

00	N/A (less than 3 years old)	10	High school graduate
01	No school completed	11	Some college, but no degree
02	Nursery school	12	Associate degree in college, occupational program
03	Kindergarten	13	Associate degree in college, academic program
04	1st, 2nd, 3rd, or 4th grade	14	Bachelor's degree
05	5th, 6th, 7th, or 8th grade	15	Master's degree
06	9th grade	16	Professional degree
07	10th grade	17	Doctorate degree
08	11th grade		
09	12th grade, no diploma		

From the information about the meaning of the codes, I made up the table shown to the right. I put this table to the right of all the data, selected all of it below the column headings, and named that range SchoolYears. The YearsSch column in the data was in column AK in the spreadsheet, so in cell F2 I put @VLOOKUP(AK2,$SchoolYears,1) and copied this to all the cells below it in column F. The variable made up in this way approximates the years of education and will be called Education.

While it is clear that some basic education has a positive effect on earnings, we may still wonder whether there are not perhaps decreasing returns to education. To allow for that possibility, we can put into column G the square of Education.

Next, into column H let us put the Sex column of the data. A 0 indicates male; and a 1, female. Into column I we can put the Age variable.

With these variables in place, we can now do a regression. I suggest that you put the results below the data and lined up so that the regression coefficient for a column of the X array is directly below that column. My results, calculated for sample 0, are shown in the first column of Table 7.3 below.

Exercise 7.4: Perform similar calculations for your sample. Compare your results with those for sample 0. Do not be dismayed is they are slightly different, for your sample is different.

You will notice that your spreadsheet program computes ‟standard deviations” of the regression coefficients. The regression coefficients from a sample , just like the mean of a variable in a sample, are random variables, so it is not surprising that they should have standard deviations. How those standard deviations may be computed and under what assumptions those computations are valid, however, is not immediately obvious. We shall come to that question later in this chapter; here we concentrate on the regression coefficients.

We put the square of Education into the regression to see whether perhaps there were diminishing returns to education, as would be indicated by a negative sign on this variable. On the contrary, we found a positive sign on this variable and a *negative* sign on Education itself. This negative sign may at first seem puzzling; to resolve the puzzle, let us write the Education effect as the sum of the two terms involving education:

$$\text{Education effect} = -1687*\text{Education} + 188.7*\text{Education}^2.$$

The marginal effect of an additional year of education is then

$$\text{Marginal effect of additional education} = -1687 + 2*188.7*\text{Education}.$$

At five years of education, the marginal effect of an additional year of education on annual earned income is $200 per year; at eight years, $1332; at 12 years, $2842; at 16 years, $5861, so the results are definitely consistent with the idea that education influences earning ability.

We failed to find decreasing returns for education, but what about for age? Perhaps few people get older just so they can earn more, but let us nonetheless put into the regression the square of age. The results are shown in the second pair of columns in Table 7.3. The coefficient on age has gone up from 321.8 with a standard deviation of 44.7 to 2044.2 and a negative coefficient, -20.7, has appeared on the square of age. Indeed, there are diminishing returns to age; and you can quickly calculate that, financially speaking, it is not a good idea to let your age advance beyond about 50.

A lesson from this story which you can more easily apply, however, is that *the standard deviation of a regression coefficient gives you no idea whatsoever about what will happen to that coefficient when an additional variable is added to the regression.* We have just seen the coefficient on age increase more than four standard deviations when another variable was added.

	Educ. sex, age		+age squared		+Division		+Hours & HoursSq		Reduced Sample	
	RegCoef	*StdErr*	RegCoef	*StdErr*	RegCoef	*StdErr*	RegCoef	*StdErr*	RegCoef	*StdErr*
Intercept	3062.4	*6699.1*	-27473.6	*7672.5*	-25240.6	*7802.1*	-21880.0	*7613.7*	-32488.4	*9140.9*
YearsSch	-1687.2	*979.4*	-1374.6	*958.7*	-1361.7	*961.3*	-1136.8	*937.5*	-1050.0	*1062.1*
YSchSq	188.7	*38.7*	158.4	*38.1*	156.8	*38.2*	139.2	*37.3*	140.5	*41.7*
Sex	-12099.6	*1234.6*	-12439.7	*1208.2*	-12414.2	*1208.8*	-10621.3	*1197.3*	-11079.5	*1376.7*
Age	321.8	*44.7*	2044.2	*229.5*	2025.5	*230.1*	1399.9	*239.9*	1410.4	*314.0*
AgeSq			-20.4	*2.7*	-20.2	*2.7*	-13.1	*2.8*	-12.9	*3.7*
NewEng					2315.0	*2875.2*	3372.5	*2804.3*	4056.1	*3287.1*
MidAtlantic					167.0	*2168.5*	-29.5	*2113.1*	964.4	*2416.3*
ENCentral					-3422.8	*2141.7*	-3413.5	*2086.5*	-2852.3	*2410.7*
WNCentral					-874.5	*2739.7*	-1699.7	*2673.9*	-761.2	*3078.8*
SAtlantic					-3048.7	*2105.7*	-3344.4	*2052.2*	-3329.7	*2342.4*
ESCentral					183.0	*2889.2*	-509.2	*2816.1*	-570.5	*3191.8*
WSCentral					-5497.1	*2397.0*	-5901.7	*2335.8*	-5906.0	*2674.5*
Mountain					-3098.8	*2901.7*	-3645.8	*2830.0*	-2841.5	*3230.6*
Hours							193.2	*79.0*	548.0	*159.7*
HoursSq							1.3	*1.2*	-2.2	*1.8*
RSq	0.2277		0.2620		0.2696		0.3078		0.2982	

Table 7.3: Regressions Explaining Earned Income

Exercise 7.5: Add age squared to the regression for your sample. Which coefficients changed a lot? How do your results compare with those for sample 0?

Dummy variables

A next step in our study of earned income can well be to ask Are there regional differences? Would a person with the same age, education, and sex expect to earn more in New England or on the Pacific Coast? To study that question, we can use the data on the geographic division of the country in which the person lives. This variable, called Division, has the value 1 if the person lives in New England, 2 if in the Mid Atlantic, 3 if in East North Central, and so on through 9 for someone living on the Pacific coast. Clearly, we cannot just throw this variable into the regression, for there is no reason to think that the effect of living in these different regions on earned income increases linearly with the number of the region. No, we have to find a different way. We will make up from this single variable a set of *dummy variables*, variables that are always just 0 or 1. The first will be 1 if the person lives in New England and otherwise 0; the second will be 1 if the person lives in the Mid Atlantic region and otherwise 0, and so on for all 9 regions. The regression coefficient on each of these variables tells how much a person living in that region earns more or less than an individual in the basic reference region. But wait! If we put in variables for all nine regions, there will be no reference region! (Mathematically, the intercept column of the X array is a linear combination of the dummies; one of the pivot elements in the solution of the normal equations will be zero, so the solution will be indeterminate.) Consequently, the dummy for one region must be left out. I have left out the Pacific region, so that it becomes the reference region. Thus, the coefficient on the New England dummy will show how much more or less a person living in New England would earn than would a person of the same sex, age, and education living on the Pacific coast.

Exactly how to make up dummy variable depends slightly on the spreadsheet you are using. In 1-2-3 release 5, I use the @IF function. Specifically, the Division variable is in column AY, so in the top cell of the New England dummy's column (which is in row 2) I put: @IF($AY2=1,1,0) . This function looks at the value in the AY2 cell and if it is 1 (the code for New England), it puts a 1 in this cell; otherwise it puts a 0. Next, copy this cell across the row to the other dummy columns and edit them by replacing the 1 on the right of the = by 2, 3, etc. Then copy this row of dummies to all the other individuals. Finally, we repeat the regression including the new variables. The results are shown in the third pair of columns of Table 7.3. In sample 0, New England emerged as the high-earning area, with earnings of $2315 above the Pacific region. It was followed by the East South Central, Mid Atlantic, Pacific, and on down to West South Central. However, the computed standard errors were high on all of these variables so there is reason to expect that you may get very different results for your sample.

Exercise 7.6: Add regional dummies to the regression for your sample. How do your results compare with those for sample 0?

Proxy variables

So far, our regression has taken no account of the effort the person makes to earn money. Yet such effort is certainly likely to affect earnings. One would like something like hours worked per year or hours worked in a typical month or week. The only variable anything like that, however, is "hours worked last week," — that is, the week before the Census was taken. The question is probably formulated that way in hopes of getting a more accurate response than would a question about the average or typical workweek. Nonetheless, looking over the sample shows a number of people with zero "hours worked last week" yet earned incomes of over $20,000. They may have been unemployed "last week," or on vacation, or on sick leave, or have seasonal employment. Thus, this variable is not what we really want, but it is the best available measure of effort. We know from our experience with "Age Squared" that the coefficients of variables in the regression may be strongly influenced by omitting variables which play a role in determination of the dependent variable. So what are we to do? The usual procedure is to bemoan the quality of the data, through up one's hands, and use the *proxy* variable, the nearest thing we can get to the variable we would like to have. ("Proxy" is from Latin *proximus*, nearest, next, most akin, most like.) We shall follow that procedure. So we put in Hours and Hours squared and declare that they are proxies for effort.

The results are shown in the fourth pair of columns. The coefficient of about $193 per year for an hour worked per week works out to about $4 per hour for the first hour worked. For someone working 40 hours per week, however, the value of the marginal hour worked is up to about $6. One important effect of including the effort proxy is that the large negative coefficient on the sex variable is reduced somewhat, from - $12,414 to -$10,621, still a large number.

With people who report 7 hours of work during the previous week and an annual earned income of $4000, one may well imagine that the reported work week is a typical. But where there are earnings of $20,000 with zero hours worked, one suspects that the proxy is really pretty bad. A good bit of ingenuity has been expended on the question of what to do in this case. One possibility (not without its own problems) is to restrict the regression to the individuals who report positive hours worked. We can easily do that by sorting the sample on Hours and excluding from the regression the individuals who reported 0 hours.

The result is shown in the last pair of columns in Table 7.3. The coefficient of $548 for the first hour per week works out to about $11 an hour for the first hour per week, but this time the coefficient on the Hours squared variable is negative so that the value of a marginal hour for one working 40 hours is only about $7.50. These higher values with the reduced sample probably mean that Hours is a better proxy for effort for this sample.

Note that reducing the sample changed some of the regional coefficients quite a lot. This result is in line with their large computed standard errors, which indicate that they may be quite sensitive to changes in the sample.

Exercise 7.6: Add the hours proxy for effort to the regression for your sample. How do your results compare with those for sample 0?

Logarithmic variables

It has perhaps occurred to you that the additive form of the function that we have been estimating may be inappropriate. For example, is it appropriate to assume, as the additive form does, that being a woman reduces earned income by a constant amount no matter what the woman's education, age, region, or hours worked? Might it not be more realistic to assume that her earned income is reduced by a constant fraction of what it would have been for a male with these same characteristics?

If instead of estimating

$$y = b_1 x_1 + b_2 x_2$$

we estimate

$$\ln y = b_1 x_1 + b_2 x_2$$

then if we take the partial derivative of both sides with respect to x_1 we have

$$\frac{1}{y}\frac{\partial y}{\partial x} = b_1 .$$

In other words, b_1 is the *proportional* change in *y*, not the absolute change, when x_1 changes by one unit.

Let us apply this idea to our regression. Use the @LN() function (not the @LOG() function) to put the natural logarithm of earnings in column B. Then do the regression again. The results for sample 0 are shown in the first pair of columns in Table 7.4.

The first result to notice is a striking increase in RSQ, which is now up to .48, a respectable number for this sort of work. Secondly, we must now change the interpretation of all of the coefficients. The coefficient of -.4599 on the sex variable does not mean that the woman earns 46 cents less; rather it means that the natural logarithm of her earnings is .4599 below that of her male "twin." If his earnings are M and hers are F, then

ln F = ln M - .4599

and taking the exponential funtion of both sides gives

F = M exp(-.4599) = .631 M.

In other words, her wages were 37 percent lower than his.

For coefficients close to 0, you can use the approximation

ln(1 + x) = x

to judge the percentage directly from the coefficient. For example, from the coefficient of .1161 on the New England dummy, we may quickly judge that the earnings of someone living in that region will be roughly 11.6 percent above that of a “twin” living in the Pacific region. The correct multiple is exp(.1161) = 1.1231 or 12.3 percent above the Pacific twin's earnings.

The results of using only individuals with positive hours worked is shown in the second pair of columns. As before, the principal difference is in a stronger coefficient on hours.

If we had a good measure of hours worked, it would be natural to suppose that earnings would be proportional to hours: Earnings = W*Hours, where W is something like the wage rate. But then we should have

$$\ln(\text{Earnings}) = \ln(W) + \ln(\text{Hours}),$$

In other words, we should be regressing the logarithm of Earnings not on Hours, as we have so far, but on the logarithm of Hours. Furthermore, the regression coefficient should be 1.0.

	Full Sample		Reduced Sample		Full Sample		Reduced Sample	
	RegCoef	*StdErr*	RegCoef	*StdErr*	RegCoef	*StdErr*	RegCoef	*StdErr*
Intercept	5.7098	*0.3132*	5.1997	*0.3136*	5.4910	*0.3205*	3.8576	*0.3358*
YearsSch	0.0286	*0.0386*	0.0176	*0.0364*	0.0270	*0.0395*	0.0213	*0.0367*
YSchSq	0.0026	*0.0015*	0.0028	*0.0014*	0.0028	*0.0016*	0.0026	*0.0014*
Sex	-0.4599	*0.0492*	-0.4173	*0.0472*	-0.5003	*0.0501*	-0.4276	*0.0474*
Age	0.1134	*0.0099*	0.0990	*0.0108*	0.1259	*0.0100*	0.1064	*0.0107*
AgeSq	-0.0012	*0.0001*	-0.0010	*0.0001*	-0.0013	*0.0001*	-0.0011	*0.0001*
NewEng	0.1161	*0.1154*	0.0782	*0.1128*	0.0926	*0.1181*	0.0785	*0.1136*
MidAtlantic	-0.0354	*0.0869*	0.0141	*0.0829*	-0.0340	*0.0891*	0.0240	*0.0835*
ENCentral	-0.2238	*0.0858*	-0.2040	*0.0827*	-0.2251	*0.0880*	-0.1824	*0.0834*
WNCentral	-0.2506	*0.1100*	-0.1798	*0.1056*	-0.2399	*0.1125*	-0.2020	*0.1062*
SAtlantic	-0.1221	*0.0844*	-0.1629	*0.0804*	-0.1230	*0.0865*	-0.1583	*0.0810*
ESCentral	-0.1934	*0.1158*	-0.2047	*0.1095*	-0.1828	*0.1187*	-0.2018	*0.1104*
WSCentral	-0.2363	*0.0961*	-0.2895	*0.0918*	-0.2287	*0.0985*	-0.2812	*0.0925*
Mountain	-0.2285	*0.1164*	-0.1478	*0.1108*	-0.2276	*0.1192*	-0.1647	*0.1115*
Hours	0.0351	*0.0032*	0.0738	*0.0055*				
HoursSq	-0.0002	*0.0000*	-0.0006	*0.0001*				
LnHours					0.2717	*0.0188*	0.8590	*0.0535*
RSq	0.4824		0.4969		0.4558		0.4884	

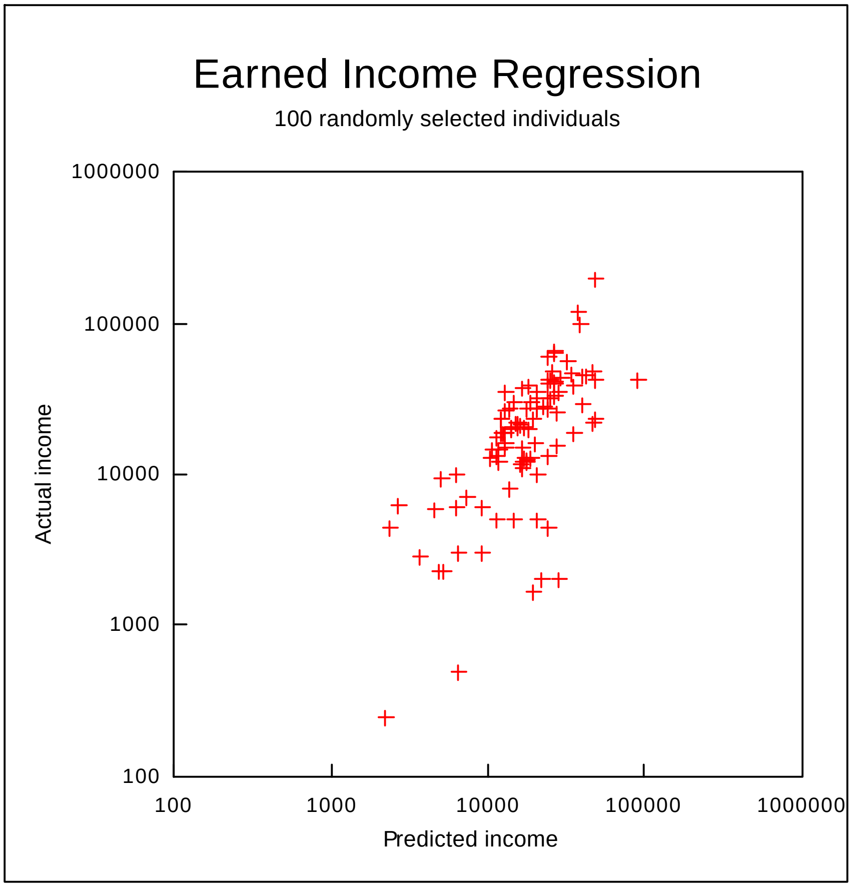
Table 7.4 Regression of Logarithm of Earned Income

But before we rush off to replace Hours with ln(Hours), we had best recall that the logarithm of 0 is minus infinity. So now we really must do something special about the observations with 0 hours. I have tried two possibilities. In the first, I simply put a 0 in the ln(Hours) column if Hours were 0. The results are shown in the third pair of columns of Table 7.4. It is disconcerting that the coefficient on ln(Hours) is only .2117. The second possibility was to exclude these observations. Those results are shown in the last pair of columns in Table 7.4. Here the coefficient on ln(Hours) is .859, reasonably close to our expected value of 1. The "female discount" is slightly reduced, to -.4276, which works out to a 35 percent "discount."

Exercise 7.7: Convert the dependent variable and the hours variable to logarithms and repeat your regression on the observations with positive hours worked.

Graphing the results

Perhaps you are curious to see in a more graphical fashion how well the independent variables succeed in explaining the dependent variable. The figure below shows the predicted earnings and actual earnings for 100 randomly selected individuals with positive hours worked. The regression used was the last pair of columns in Table 7.4. The spreadsheet's ability to multiply matrices was used to compute the predicted values, which were then converted from logarithms back to dollars. The part of the sample with positive hours worked was then put in random order in the usual way, and the graph was drawn for the first 100 individuals. Both axes of the graph use a logarithmic scale. As more points are put on the graph, many of them fall on top of others near the center and become lost to the eye. Since only the outliers show up, a graph like this with 200 or 300 points tends to exaggerate the appearance of scatter.


Earned Income Regression
100 randomly selected individuals
1000000
100000
10000
1000
100
Actual income
100
1000
10000
100000
1000000
Predicted income

Chapter 8

Standard Deviations, Loss Limits, and Mexvals

We have already noticed that, when the spreadsheet programs compute regression coefficients, they also compute numbers which purport to be the standard errors of these coefficients. In this chapter, we will see how those numbers are computed and under what conditions they can be thought of as standard deviations of the regression coefficients. We will also show how to derive from them other, more strictly descriptive statistics which are always valid.

All of these discussions will be greatly facilitated by the use of matrix notation, which will be introduced and applied to our problem in the next section.

1. Matrix notation for regression

To introduce matrices, let us consider again the equations we just solved. They were

$$\begin{aligned} 5b_1 + \ 25b_2 + \ 25b_3 &= \ 75 \\ 25b_1 + 225b_2 + \ 85b_3 &= 380 \\ 25b_1 + \ 85b_2 + 171b_3 &= 415\ . \end{aligned}$$

We could economize on b's and + signs if we would write them instead as

$$\begin{pmatrix} 5 & 25 & 25 \\ 25 & 225 & 85 \\ 25 & 85 & 171 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 75 \\ 380 \\ 415 \end{pmatrix}$$

We have changed nothing by this rewriting; we have just introduced a shorthand. Now let us think of the array of numbers on the left as a single entity and call it A. This A is a *matrix*. In fact, any rectangular array of numbers is a matrix. The column of numbers on the right side is then also a matrix, but since it has only one column it may also be called a *vector*. Because it is a vector, we will denote it with a lower case letter and call it *c*. The column of the unknown b's is also a vector, and we may as well call it *b*. Then the whole equation can be written as just

$$Ab = c.$$

In this equation, we say that the matrix *A* is "post multiplied" by the column vector *b*. What that means is fully explained by looking back at the original equations. Here are a couple of examples to check your understanding of multiplying a matrix by a vector:

$$\begin{pmatrix} 3 & 7 \\ 5 & 2 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 23 \\ 19 \\ 14 \end{pmatrix} \qquad \begin{pmatrix} 3 & 1 & 2 \\ 4 & 0 & 8 \\ 2 & 1 & 5 \end{pmatrix}\begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 19 \\ 28 \\ 17 \end{pmatrix}.$$

Note that in order to post multiply the matrix *A* by the column vector *b*, *b* must have the same number of rows that *A* has columns. The result will be a column with as many rows as A has rows. The number of rows and columns of a matrix are called its dimensions. The matrix on the left in the first example above is said to be "3 by 2", that is it has 3 rows and 2 columns. In general, we write "*A* is (m,n)" when we mean that *A* has *m* rows and *n* columns.

Now suppose that we have two matrices, *A* and *B*. If they are of the same dimensions, we can define their sum, $A + B$, as the matrix composed of the sums of the corresponding elements in the two matrices. For example,

$$\begin{pmatrix} 3 & 5 \\ 4 & 3 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}.$$

When it comes to the product of two matrices, *AB*, however, we do not define it as the matrix composed of the products of corresponding elements. Rather we define it as the matrix whose first column is the product of *A* post-multiplied by the first column of *B*, and whose second column is the product of *A* post-multiplied by the second column of *B*, and so on. Here are two examples:

$$\begin{pmatrix} 3 & 5 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} 21 & 28 \\ 17 & 19 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 2 & 1 \\ 2 & 5 & 3 \end{pmatrix}\begin{pmatrix} 5 & 2 \\ 1 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 19 & 13 \\ 21 & 22 \end{pmatrix}.$$

In order for the product *AB* to be defined, *B* must have as many rows as *A* has columns. The product will have as many rows as does *A* and as many columns as does *B*. In general *AB* will not be the same as *BA*, though there are important special cases in which they are the same.

It is easy to verify that $(A + B) + C = A + (B + C)$ –the order in which we add matrices makes no difference – and that $(AB)C = A(BC)$ – the order in which we multiply makes no difference. Also, as with ordinary numbers, multiplication is distributive over addition: $A(B + C) = AB + BC$.

To express our least-squares problem in terms of matrices, we need just one more concept, the *transpose* of a matrix. The transpose of *A*, denoted by *A'*, is the matrix whose first column is the first row of *A*, whose second column is the second row of *A*, and so on. For example, if *A* is the matrix

$$A = \begin{pmatrix} 5 & 2 \\ 1 & 3 \\ 2 & 1 \end{pmatrix}$$

then *A'* is

$$A' = \begin{pmatrix} 5 & 1 & 2 \\ 2 & 3 & 1 \end{pmatrix}.$$

If *A* = *A'*, then the matrix must be square and is said to be *symmetric*. You can, with a little thought, see that *(AB)'* = *B'A'*.

Now let us denote by *X* the matrix of observations on the independent variables and let y denote the vector of observations on the dependent variable in our previous regression example. Thus,

$$X = \begin{pmatrix} 1 & 10 & 5 \\ 1 & 5 & 1 \\ 1 & 0 & 6 \\ 1 & 10 & 3 \\ 1 & 0 & 10 \end{pmatrix} \qquad y = \begin{pmatrix} 17 \\ 10 \\ 12 \\ 16 \\ 20 \end{pmatrix}$$

Finally let b denote the vector of regression coefficients. We want to choose b to minimize S, the sum of squared residuals. In matrix notation, S is

$$S = (y - Xb)'(y - Xb).$$

The minimizing b is given by the equation

$$(X'X)b = X'y\,. \tag{8.1.1}$$

It may take a moment to see that this equation is the same as the previous equation 10. But just write out *X'* and *X* and start forming the product *X'X*, and you will soon realize that you are forming the matrix A with which we began this section. Note that *X'X* is symmetric.

How can we show the process of solution of linear equations in matrix notation? To do so we first need a special notation for any square matrix that has 1's on the diagonal running from top left to bottom right and is otherwise zero. Such a matrix is called an "identity" matrix and is therefore denoted by *I*. Note that for any matrix *M*, *IM = M* and *MI = M*, where *I* is an identity matrix of appropriate dimension. Suppose now that we have the matrix equations *Ax = c*, where *A* is a square matrix and *x* and *c* are vectors. If we can find some matrix *B* such that *AB = I*, then *A(Bc) = (AB)c = Ic = c*, so *x = Bc* is the solution of the equations *Ax = c*.

But how do we find the matrix *B*? Since $AB_1 = I_1$, $AB_2 = I_2$, and $AB_3 = I_3$, where the subscript denotes a column of *B* or *I*, we could just solve these equations one-by-one for the columns of *B*. There is, however, an easier way; we can solve them all at once. Table 8.1 shows how. In the first three rows and first four columns of each panel you will recognize the corresponding panel of Table 2.1. In the fifth column we have written the right-hand side of the equations $AB_1 = I_1$.

5.	25.	25.	75.	1.	0.	0.
25.	225.	85.	380.	0.	1.	0.
25.	85.	171.	415.	0.	0.	1
75.	380.	415.	1189.	0.	0.	0.
1.	5.	5.	15.	0.2	0	0.
0.	100.	-40.	5.	-5.	1.	0.
0.	-40.	46.	40.	-5.	0.	1.
0.	5.	40.	64.	-15.	0.	0.
1.	0.	7.	14.75	0.45	-0.05	0.
0.	1.	-0.4	0.05	-0.05	0.01	0.
0.	0.	30.	42.	-7.0	0.4	1.
0.	0.	42.	63.75	-14.75	-0.05	0.
1.	0.	0.	4.95	2.083	-0.143	-0.233
0.	1.	0.	0.61	-0.143	0.015	0.013
0.	0.	1.	1.4	-0.233	0.013	0.033
0.	0.	0.	4.95	-4.950	-0.610	-1.400

Table 8.1: Regression with Matrix Inversion

Now notice that if we carry the pivot operations through this column as well as the first four we will have in it in the fourth panel the solution of the equation $AB_1 = I_1$. Note that appending this extra column had absolutely no effect on the previous four columns. Nor did the fourth column, the original right-hand side of the equations, have any effect on what happened in the fifth column. We can, therefore, append the other columns of *I* as the last two columns of the first panel in Table 8.1, carry through the pivoting on them, and get in the top three rows of the last panel the matrix *B*. We have gotten the solution for three sets of equations for considerably less than three times the work required for one.

The matrix *B* which we have found in this way is called the "inverse" of *A* and is denoted by A^{-1}. From its construction, we know that $AA^{-1} = I$. It is also true that $A^{-1}A = I$, for since

$$AA^{-1} = I$$

pre-multiplying by A^{-1} gives

$$A^{-1}AA^{-1} = A^{-1}$$

and post-multiplying by the inverse of A^{-1}, $(A^{-1})^{-1}$, gives

$$A^{-1}AA^{-1}(A^{-1})^{-1} = A^{-1}(A^{-1})^{-1} = I$$

or

$$A^{-1}A = I.$$

Thus, every right inverse is also a left inverse.

We will also often use the fact that the inverse of a symmetric matrix is itself symmetric. If $A = A'$, then

$$I = (AA^{-1})' = (A^{-1})'A' = (A^{-1})'A$$

and post multiplying both sides of this equation by A^{-1} gives

$$A^{-1} = (A^{-1})',$$

which is to say, A^{-1} is symmetric.

We can now summarize our discussion so far in one equation: the regression coefficients are given by the equation

$$b = (X'X)^{-1}X'y. \tag{8.1.2}$$

For future reference, we need one more concept, the *trace* of a square matrix. The trace is simply the sum of the diagonal elements. Thus, if C is (m,m), then the trace of C, $tr\ C$, is just

$$tr\ C = \sum_{i=1}^{m} c_{i\,i}\,.$$

If A is (m,n) and B is (n,m) then AB and BA are both defined but are not generally equal to one another or even of the same dimension. But both are square so that tr(AB) and tr(BA) are both defined. Now it is a remarkable and useful theorem that tr(AB) = tr(BA). To see why this is true, just note that any element of A, say $a_{1\,3}$, will enter into the sum that forms tr(AB) exactly once and that one time it will be multiplied by the symmetrically placed element of B, $b_{3\,1}$. Now notice that the element also enter tr(BA) exactly once and is again multiplied by the same element of B. Thus the sums that form tr(AB) and tr(BA) consist of exactly the same elements and must therefore be equal to one another. Working a small example will make this proof plain.

Exercise 8.1. For the data of exercise 7.1, compute $(X'X)^{-1}$, then $X'y$, and then, by matrix multiplication, $b = (X'X)^{-1}X'y$.

3. Standard deviations of regression coefficients

Regression coefficients are generalizations of the idea of the mean. Indeed, the regression of a variable, *y*, on just a vector of 1's, is just the mean of *y*, as you can quickly verify. Does it make sense to talk about the standard deviation of a mean? Maybe yes; maybe no. The grades on the mid-term examination in this course this semester have a mean. Does it make sense to talk about the standard deviation of that mean? For practical purposes, no. You can, of course, imagine a vast sea of students who might have taken it but did not and never will, for exactly the same course and exam will never be given again. If, at the end of the course, I do a regression of the final grade on the midterm grade, the regression coefficient will be descriptive of this class. Only by invoking the dubious hypothesis of the vast sea of totally non-existent students is it possible to talk about the standard deviations of that regression coefficient.

Only if we are speaking of the mean of variables in a random sample drawn from a much, much larger population does it make much sense to talk about the standard deviation of the sample mean. The same is true of regression coefficients. *The standard deviations of the regression coefficients make little sense if the regression is done on the whole population.* I must stress this elementary point because it is often forgotten. The regression programs do not know whether they are working with a sample or the whole population, so they spit out numbers labeled "standard deviations" even when working with the whole population. This point becomes especially important when we come to regression on time series data. We may certainly regress GDP on M2 money supply in the 1980's. The regression coefficient that results may be a most interesting descriptive statistic, but it has no *meaningful* standard deviation, because the regression has used all the data. Do not say that we could go back to the 1970's or forward to the 1990's. The economy was quite different in those decades; the 1980's in no sense represent a *random* sample from, say, the 20th century.

We will derive standard deviations of regression coefficients under the *assumption* that the *y* vector which we observe was generated from the matrix *X* of independent variables by the equation

$$y = X\beta + \epsilon \qquad (8.3.1)$$

where β is a constant vector and ϵ is a random variable with zero mean independently drawn from the same identical distribution for all observations.

There are two different interpretations of this equation; one that I may call the earthly and the other the heavenly. In the earthly interpretation, β is simply the *b* which would be found if the regression were done on the entire population, not just the sample. In this case, there is no question about the existence of β; there is, however, considerable question about whether or not it is possible to maintain that the residuals from that equation could be thought of as independent drawings from the same identical distribution. For example, it might turn out in our example in the previous chapter that the standard deviation for men was greater than for women.

The alternative interpretation, which I call the heavenly, is that there is a great Datamaker who takes the true β and the X matrix, draws a random vector ϵ, computes *y* by (8.3.1), bundles y and X together in neat bundles and throws them out, one after another, into the cosmos. One hit the earth, exploded, and created the data which you are using. Others were caught elsewhere, but not on this earth. Everywhere one is caught, someone computes the *b* and sends in the estimate to the cosmic data center where mortals like you and me who do not know the true value of β take the mean and the standard deviations of the *b*'s that pour in. Under this interpretation, it is always meaningful to talk about the standard deviation of the *b* vectors even though you will never see more than one of them. There is, however, a problem. You may be sure that you know the true *y* and *X*, but Datamaker sometimes throws jokers into the *X* matrix, that is, variables which had coefficients of 0 in β. The researcher's job, in this interpretation, is to compute confidence intervals for the regression coefficients and to point out the variables that may be jokers because their confidence intervals include 0.

You may be surprised to learn that the heavenly interpretation is virtually the only one ever taught. The mathematics for determining variances and covariances of the regression coefficients is not affected by the distinction between the assumptions, but how one treats the results is much affected.

To develop the formulas for the variances of the regression coefficients, we will find it convenient to use the term *expected value* of a function of a random variable. If *x* is a random variable with density function f(x) and *g(x)* is a function of *x*, the *expected value* of *g(x), E(g(x)),* is

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f(x) dx .$$

We are already familiar with

$$E(x) = \mu \quad and \quad E((x - \mu)^2) = \sigma^2 .$$

If *c* is a constant, clearly *E(cx) = cE(x).*

Using the expected value notation, we can express our assumption about ϵ by

$$E(\epsilon) = 0 \quad and \quad E(\epsilon'\epsilon) = \sigma^2 I. \tag{8.3.2}$$

where 0 denotes a vector of zeroes.

With these preliminaries out of the way, we can get down to business. What are the properties of the least squares regression coefficients, *b*? Let us look first at their expected value:

$$E(b) = E((X'X)^{-1}X'y) = E\big((X'X)^{-1}X'(X\beta + \epsilon)\big) = \beta + (X'X)^{-1}X'E(\epsilon) = \beta. \tag{8.3.3}$$

The first equality here used equation (8.1.2); the second used (8.3.1); the third just multiplied out the expression and used the definition of the inverse and the assumption that the *X* matrix is constant, non-random, and the last then used (8.3.2). This result is often expressed by saying that *b* is an *unbiased* estimate of β.

The matrix of variances and covariances of the regression coefficients is

$$
\begin{aligned}
E(b - \beta)(b - \beta)' &= E((X'X)^{-1} X'e\, e'X(X'X)^{-1}) \\
&= (X'X)^{-1} \sigma^2 I\, X(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1}. \qquad (8.3.4)
\end{aligned}
$$

so the variances of the b's are the diagonals of this matrix; the standard deviations are their square roots. If we knew σ^2, we could calculate the standard deviations precisely. In fact, we never know σ^2 and must estimate it. The most natural estimate might be r'r/T, the variance of the residuals. This estimate would be biased, for – as we shall show –

$$E(r'r) = (T - n)\sigma^2,$$

where T is the number of observations or rows of X and n is the number of independent variables, or columns of X. To see why this formula holds, note first that

$$
\begin{aligned}
r &= y - Xb = X\beta + e - X(X'X)^{-1}X'(X\beta + e) \\
&= e - Me
\end{aligned}
$$

where $M = X(X'X)^{-1}X'$. This M is a remarkable matrix. Note that M = M', and M'M = MM = M and that

$$\text{tr } M = \text{tr } (X(X'X)^{-1})X' = \text{tr } X'X(X'X)^{-1} = \text{tr } I = n,$$

where I is the (n,n) identity matrix. Now

$$r'r = (e - Me)'(e - Me) = e'e - 2e'Me + e'M'Me = e'e - e'Me.$$

Since r'r is (1,1), r'r = tr r'r. So

$$
\begin{aligned}
E(r'r) &= E(\text{tr } r'r) = E(\text{tr}(e'e - e'Me)) = E(e'e) - E(\text{tr}(ee'M)) \\
&= T\sigma^2 - \text{tr}(E(ee'M)) \qquad \text{(Since expected value of a sum is the sum of the expected values.)} \\
&= T\sigma^2 - \text{tr}(\sigma^2 IM) \qquad \text{(Where I is T by T)} \\
&= T\sigma^2 - \sigma^2(\text{tr } M) = (T - n)\sigma^2.
\end{aligned}
$$

Thus, if we use $s^2 = r'r/(T - n)$, we will have an unbiased estimate in the sense that $E(s^2) = \sigma^2$.

The "standard deviations" or "standard errors" which your spreadsheet or regression program gives are calculated from (8.2.4) with this estimate of σ^2.

$$
\begin{aligned}
s_1 &= \text{sqrt}[(4.95/2)*2.083] = 2.27 \\
s_2 &= \text{sqrt}[(4.95/2)*.015] = 0.19 \\
s_3 &= \text{sqrt}[4.95/2)*.033] = 0.29
\end{aligned}
$$

Exercise 8.2. Calculate the standard deviations of the regression coefficients for the data in exercise 7.1. Compare with the output from regression done with your spreadsheet.

These standard deviations can be used, just as were the standard deviations of the mean, to set up confidence intervals for the coefficients which would be found if we could do this regression for the whole population. If the error terms are normal, the regression coefficients will be normal. Even if the error terms are not normal, random sampling and the central limit theorem can usually be relied upon to produce approximately normally distributed regression coefficients as the sample size increases beyond, say, fifty.

All of the above discussion has assumed that the *X* matrix was fixed, not random. But when we do sampling, the *X* matrix will be different for each sample. Does that fact complicate matters? Somewhat, but not a lot. E(b) is still β. The variance-covariance matrix of *b*, however, is different for each sample. If you are lucky, your *X* matrix will give narrow confidence intervals; if it is not your day, your confidence intervals will be broader. But if you make 95 percent confidence intervals on a large number of random samples, they will include almost certainly include the population value of the regression coefficient in about 95 percent of the cases.

If, however, we add a variable to the regression, all bets are off. The new coefficient may be far outside the 99 percent confidence interval around the coefficient first estimated. In fact, we almost never know exactly what variables should be included in *X*. In fact, the appropriate variables for the regression may not be in our data set. Thus, *the standard deviations give us a very limited confidence in our knowledge of the value of the regression coefficient. Further, if we have done a regression over the whole population of interest, they are nearly meaningless.*

The notion of adding a variable wreaks havoc with the heavenly interpretation of (8.3.1). If you didn't have all the necessary variables in the *X* matrix, then your b was not only not an estimate of the full β, it wasn't even an unbiased estimate of the elements of β corresponding to elements of *X* you did have. The earthly interpretation suffers from no such problem. When you added a variable to X, you changed the values of the β vector for the whole population. Both *b* vectors that you estimated from your sample were unbiased estimates of the β that would be found from the corresponding regression on the whole population values.

4. Tests of "significance"

Most courses in statistics give much attention to tests of "significance." Briefly, a regression coefficient is said to be "significant at the 5 percent level" if the 95 percent confidence interval for it does not include 0. Two types of errors are then noted. In the Type 1 error, a coefficient is accepted as non-zero when it is in fact zero; in the Type 2 error, a coefficient is declared insignificant when in fact it is non-zero.

This is a very peculiar use of the word "significant." Consider two cases. In the first, we find that a woman has $10,000 per year lower earned income that her male "twin" and that the 95 percent confidence interval is from -$1,000 to +$21,000. In the second case, we find that this regression coefficient is $5 with a confidence interval from $1 to $9. In which case have we found a significant earning difference? I would say that in the second case we can be certain that the earnings differential is utterly insignificant, not worth bothering about, while in the first case it is highly probable that the differential is very important. But in standard statistical testing parlance, the second case is "significant" while the first is "insignificant."

I urge you to avoid this perverse use of language. Think carefully about the meaning of each regression coefficient, as we have tried to do in the example. Base any observations about the importance of the coefficient on an economic interpretation of its meaning. Say, if you must, that its 95 confidence interval contains 0, so that what appears important may be just an accident of sampling. If you are working with the whole population, don't even mention the confidence intervals, because they are virtually meaningless.

It is common practice among statistical workers to try a number of explanatory variables and then to eliminate those whose coefficients are less than, say, twice their standard deviation. Thus, the reported results all look "significant." Under the "heavenly" interpretation, however, the distributions of the regression coefficients found in this way are unknown and the reported "tests of significance" invalid. Why? Because in throwing out variables that did not pass the "significance" test, you may well have thrown out variables that, in fact, had a non-zero value in β. Putting that variable back in might well cause the coefficients of other variables, as we know, to jump well outside their calculated confidence intervals. The earthly assumption is kinder to this practice. Because we are not claiming to estimate anything more than the regression coefficients which we would get if we had the whole population, our claim to unbiased estimates seems fairly sound.

In fact, we almost never know *a priori* which variables to put into X. If we put in every variable in sight, we are apt to get many nonsense coefficients. If we drop out something that belongs in, we bias (under the heavenly interpretation) other coefficients. There is no mechanical procedure for solving this problem. My advice is to pay close attention to the economic meaning of the coefficients. Include variables that have coefficients of reasonable value and contribute to the fit; exclude those that contribute little to the fit or have implausible values for coefficients. But don't claim that the results are "statistically significant" because they are twice the standard deviations produced by the regression program.

Exercise 8.3: Develop an equation to explain the value of the house in which a person lives.

5. A shortcut to the sum of squared residuals

The last row of each panel of Table 8.1 provides a short-cut to the calculation of S, the sum of squared residuals. We have not yet explained this row. In the top panel, it contains the row vector (y'X, y'y), which is the transpose of the fourth column, so the first four rows and columns form a symmetric matrix. As we generate the successive panels, we carry through the pivot operations on the last row just as on the other rows.

Now let me point out a surprising "coincidence". In the fourth panel of Table 8.1 we find, in the position where y'y was originally, the number 4.95. This is exactly the value that we got for S by a totally different procedure in Table 7.2. Can it be that we can find the sum of squares by pivoting instead of by calculating all of the residuals, squaring them and summing? Yes, that is true. And it is useful in many ways. Let us see why it is so.

By the time we reach the last panel in Table 8.1, we will have subtracted from the original last row some combination of the rows above it and gotten, in the first three positions, zeroes. What combination of the rows above it did we subtract? Since we originally had in those positions y'X and, after subtracting we have 0, we must have subtracted a combination, given by the row vector c, such that c(X'X) = y'X. In fact, this c is really just b', the transpose of the vector of regression coefficients, for

$$(X'X)b = X'y$$

so

$$b'(X'X)' = y'X$$

and

$$b'(X'X) = y'X$$

since (X'X)' = X'X. Therefore what has been subtracted from the final position of this last row is b'X'y. What was originally in it was y'y, so what is left is y'y - b'(X'y). The direct approach to calculating S first calculates

$$r = y - Xb = y - X(X'X)^{-1}X'y$$

and then forms

$$\begin{aligned} S &= r'r \\ &= y'y - y'X(X'X)^{-1}X'y - y'X(X'X)^{-1}X'y + y'X(X'X)^{-1}X'X(X'X)^{-1}X'y \\ &= y'y - y'X(X'X)^{-1}X'y = y'y - b'X'y, \end{aligned}$$

which is exactly what the pivoting gave.

Now suppose for a moment that we had set out to regress x_2 on x_1. We would have formed exactly the same 2-by-2 matrix that we see in the upper left corner of panel 1 of Table 8.1 and the final result would have been the 2-by-2 in the upper left corner of panel 2. The value of S for this problem would have been 100 and the regression coefficient would have been 5. Similarly, if x_3 had been regressed on x_1, the value of S would have been 46 and regression coefficient 5. (Because x_1 is the constant 1, the regression coefficients are the means, and the values of S are the sum of squared deviations from the means.) In general, we see that in panel i+1, after i pivot operations, we can see the regression coefficients and values of S for the regression on the first i variables of each of the remaining variables. Thus, the regression panels show a great deal about the relations among the variables. Can you give an interpretation for the element in the third row and fourth column of the third panel (the number is 42)?

Please note that each pivot element was the S value for the variable about to be introduced when regressed on all the previous values. If a pivot element is zero, the regression cannot continue; but a zero pivot can occur only if the variable about to be introduced was perfectly explained by the previous variables. If this happens when, say the third variable is about to be introduced, the G program will give the message "Variable 3 is a linear combination of preceding variables," and will abort that particular regression.

A moment's study of Table 8.1 shows that in each panel three of the columns are the columns of the identity matrix. In working with a computer, these identity columns just waste space, and it is usual to store only the non-identity columns of the matrix, as shown in Figure 2.4. I will refer to this sort of table as regression with compact inversion.

Exercise 8.4. Extend your previous computation with the data of exercise 7.1 to include the "dependent variable" row in each computation. What is S when only x1 is used as an explanatory variable? When only x1 and x2 are used? When all three are used? What are the regression coefficients for x3 regressed on x1 and x2?

6. Mexvals and derivatives – measures of the importance of each variable

So far, we have not developed a way to say anything about how important any particular variable is to the whole equation. One measure designed to help in answering this question is the "mexval" of a variable. A variable's mexval, or marginal explanatory value, is defined as the percentage that SEE will increase if the variable is omitted from the regression and not replaced by any other, though the coefficients on the remaining variables are adjusted to do as well as possible without their departed comrade.

```
     5        25        25        75
    25       225        85       380
    25        85       171       415
    75       380       415      1189

   0.2         5         5        15
  -5.0       100       -40         5
  -5.0       -40        46        40
 -15.0         5        40        64

   0.45      -0.05      7.0      14.75
  -0.05       0.01     -0.4       0.05
  -7.0        0.4      30.0      42.0
 -14.75      -0.05     42.0      63.75

   2.083     -0.143   -0.233    4.95
  -0.143      0.015    0.013    0.61
  -0.233      0.013    0.033    1.40
  -4.950     -0.610   -1.400    4.95
```

Figure 8.2: Regression with Compact Inversion

Mexval is easily calculated in the process of regression with compact inversion. With n independent variables, this form of regression leads to a final panel like this:

$$\begin{array}{llll} a_{11} & \ldots & a_{1n} & a_{1m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} & a_{nm} \\ a_{m1} & \ldots & a_{mn} & a_{mm} \end{array}$$

where m = n + 1. Remember that the lower right element is the sum of squared residuals. Note also that the values of the a's do not depend upon the order in which the row reduction was done, that is, in regression terms, they do not depend on the order of introduction of the variables. Let us suppose that variable i was the last to be introduced and let us denote the elements of the panel before its introduction with a'. Then we have the following relation between the a and a':

$a_{ii} = 1/a'_{ii}$ and $a'_{ii} = 1/a_{ii}$
$a_{im} = a'_{im}/a'_{ii}$ and $a'_{im} = a_{im}/a_{ii}$
$a_{mm} = a'_{mm} - a'_{im}a'_{im}/a'_{ii}$ and $a'_{mm} = a_{mm} + a_{im}a_{im}/a_{ii}$

Thus, the drop in the sum of squares of the residuals as variable i was introduced – and the increase in that sum of squares if it is now excluded from the equation – is a_{im}^2/a_{ii}.

The standard error of estimate of the equation would therefore rise from

$$SEE = sqrt\ (a_{mm}/T)$$

to

$$SEE' = sqrt\ ((a_{mm} + a_{im}^2/a_{ii})/T),$$

where T is the number of observations. Therefore,

$$\begin{aligned} mexval_i &= 100*((SEE'/SEE)-1) \\ &= 100*(sqrt(1+(a^2_{im}/a_{ii}a_{mm}))-1). \end{aligned}$$

For the example of the text, we find

$$mexval_1 = 100*(sqrt(1 + (4.95^2/(2.083*4.95))) -1) = 83.74$$
$$mexval_2 = 100*(sqrt(1 + (.61^2/(0.015*4.95))) -1) = 143.0$$
$$mexval_3 = 100*(sqrt(1 + (1.40^2/(0.033*4.95))) -1) = 258.9 .$$

(The numbers at the right are calculated from a more precise inverse than that shown above.)

The insights of this section and the last can be combined to calculate the derivatives of regression coefficients with respect to one another. Suppose that we were to decide that we did not trust the value provided by regression for the coefficient of some variable and that we wanted to fix the value of that coefficient. What effect would that fixing have on the other coefficients? More precisely, what would be the derivatives of the other coefficients with respect to that coefficient, if the others are determined to minimize the sum of squared errors, given the value of the fixed coefficient? Suppose the original equation was written y = Xb + Zc + r, where Z is a T-by-1 vector and c is a scalar. We are asking, What is the derivative of the least-squares value of b with respect to c? The least squares estimate of b, given c, is $b = (X'X)^{-1}(X'y - X'Zc)$, from which it is clear that the derivative of the vector b with respect to the scalar c is

$$db/dc = - (X'X)^{-1}(X'Z),$$

which is just the negative of the regression coefficients of Z on all the other variables. If Z had been treated as the last independent variable, that is just the negative of what would have been in Z's column before it was pivoted on. To get it back, we just unpivot. In the notation developed in this section, if Z is in column i, we want a'_{ji}. From the pivot operation, we have $a_{ji} = 0 - a'_{ji}/a'_{ii}$ and $a_{ii} = 1/a'_{ii}$. These may be solved for $-a'_{ji} = a_{ji}/a_{ii}$. In other words, to get the derivatives of all of the regression coefficients with respect to coefficient i, we just divide the ith column of $(X'X)^{-1}$ by its diagonal element. These derivatives are very useful for seeing the sensitivity of one coefficient with respect to another. I have often seen it happen that an equation has several coefficients with nonsensical values, but by fixing one of the coefficients to a sensible value, the others also became sensible. The derivatives are useful for spotting such situations.

Exercise 8.7. Compute the mexvals for x1, x2, and x3 with the data of exercise 7.1. Compute the derivative of b_1 and b_3 with respect to b_2.

The t-values are the ratio of each regression coefficient to the estimate of its standard deviation made using this s^2. From the fourth panel of Table 8.1, we find the following t-values for the example we have developed in the text:

t_1 = 4.95/sqrt[(4.95/2)*2.083] = 2.180
t_2 = .61/sqrt[(4.95/2)*.015] = 3.131
t_3 = 1.40/sqrt[4.95/2)*.033] = 4.874.

If the e are normally distributed, and if the true value of some β_i is zero, this ratio will have a Student t distribution. (A good bit of mathematics is required to back up that simple statement; see my book *Matrix Methods in Economics* (Addison-Wesley, 1967) Chapter 6.) This distribution depends on T-n, but for values of T - n over 30, the distribution is indistinguishable from the normal. So if T-n > 30, then under all of the previous assumptions – namely the existence of a true equation of the form we are estimating, X non-stochastic, and the elements of e independent of each other but all having a normal distribution with zero mean and the same variance – we can say, "If the true value of the regression parameter is zero, the probability that we will observe a t-value of over 2.0 in absolute value is less than .05." If we observe such a value, we are then supposed to be "95 percent confident" that the true value is different from zero.

A further question is the relation of the mexvals to the t-Statistics. The t-value for the ith variable is

$t_i = a_{im}$ / sqrt($a_{mm}a_{ii}$/(T - n))

and we have seen that

mexval_i = 100 * (sqrt(1+($a^2_{im}/a_{ii}a_{mm}$))-1).

So in terms of t, the mexval for the same variable is

mexval = 100 * (sqrt(1 + t^2/(T-n)) - 1).

Thus, in the same equation where T - n is the same for all variables, t-values and mexvals convey the same information. The difference is in ease of interpretation, ease of explanation, and "honesty". The mexval is exactly what it claims to be; the t-value is an appropriate measure to look at only if one wants to claim it to be something it probably isn't. Of course, if one does – despite all experience to the contrary – believe the t-story, then one may want a t-value; and the G program provides them.

Chapter 9

Systems of Economic Statistics

In the first part of this book, we saw how to find and work with statistics that had been prepared from primary data. In the second, we learned about working with primary data, sampling, and regression on sample data. In this third part, we return to the statistics that have already been organized into systems.

1. The National Income and Product Accounts

The National Income and Product Accounts (NIPA) constitute a major accomplishment of economic theory and measurement. When we come into the world and find automobiles, computers, a theory of gravitation or of organic evolution, it is easy to suppose that they have always been there, that they are as utterly natural as stones and stars. In fact, of course, they are all human creations and represent stunning strokes of insight and much hard work. The same is true of national accounts. One tends to suppose that they just naturally appear every month like the new moon. In fact, they are perhaps the single greatest success of economic science.

The first effort to measure national income goes back to the 1650's when William Petty, a surgeon in Cromwell's army, wrote a tract with the title *Political Anatomy* to urge that England was strong enough to invade France. From church records, Petty got the number of families in each parish. Then he estimated the average income per family in each parish, and by multiplying and adding, he found the national income, which he judged to be quite large relative to the cost of a successful invasion. (Unfortunately for Petty, Cromwell died in 1658; and with him, the zeal of the English army.) While Petty's attempt was far from the sophistication of modern accounts, it had one important feature in common with them: it relied on data (the church records) that had been compiled for totally other purposes.

During World War I, the Tsar Nicholas II of Russia appointed a commission to estimate the income of Russia to ascertain whether it was feasible for Russia to continue the war against Germany. The commission, continued under Kerensky, made its report Lenin, who then made the peace of Brest-Litovsk. The commission, however, had worked out many of the principles of national accounting. Its report came into the hands of two young Russians, Simon Kuznets and Wassily Leontief, who independently made their way to the United States. During the 1920's, Kuznets, inspired by this report and now working at the National Bureau of Economic Research (a private research group then in New York City), began the construction of estimates of real gross national product of the USA. When the Depression of the 1930's left the Government desperate to do something but with little idea of even the magnitude of the problem, Kuznets was brought to Washington to set up the basis of the national accounts. By the end of the 1930's, GNP estimates were published each year, but the complete accounts could be printed on a single page.

Wassily Leontief, who reached the US only in the 1930's, began an ambitious extension of the accounts to a complete input-output table, that is to say, to a table dividing the economy into a number of industries and showing for each industry its sales to each other industry or category of final demand (in the industry's row) and its purchases from other industries and from various sorts of primary income in the columns. Such a table is shown very schematically below.

	Buyers			Final Demands					
Sellers	Agric	Mfg	Services	PersCons	Invest	Export	Import	Gov't	Total
Agriculture	0	100	0	700	50	120	-30	60	1000
Manufacturing	200	0	380	1400	500	500	-620	540	2900
Services	100	400	0	1600	150	50	-50	400	2650
Consump of Capital	100	400	300						800
Labor Income	450	1500	1400						3350
Profits	25	200	300						525
Net interest	75	200	200						475
Indirect taxes	50	100	70						220
Total	1000	2900	2650	3700	700	670	-700	1000	11920

In this made-up example, the economy has been divided into three producing industries, Agriculture, Manufacturing, and Services. The first three rows of the table (which we shall call the industry rows) describe the sales of each of these industries to each of the others and to five broad categories of final demand, Personal consumption (PersCon), Investment, Exports, Imports, and Government. The sales to any buyer includes that buyer's purchases of imported goods, so the "sale" to imports is the negative of the total value of imports of the product in the row. Thus, the sum across any industry row (excluding the total column on the right) is equal to the domestic output of the product of that industry. These sums are shown in the last column.

Below the industry rows are the value-added or income rows. The sum down any of the industry columns (excluding, of course, the total row) is equal to the total value of the product of that industry, which is just what the corresponding row summed to. Indeed, the very definition of "profit" is the total value of sales (the row sum) minus all the costs, the sum of all the entries in the column except the one in the profit row. When that difference is put into the profit row, we can be absolutely certain that the sum of every industry column equals the sum of the corresponding industry row. (Here we call "Labor income" what the accounts more completely designate as Wages and salaries + Other labor income + Proprietors' income. Similarly, what we have called "Profits" includes also what the accounts call Rental income. See Account 1 below.)

This schematic table has two rectangles enclosed entirely by double lines. The one in the upper right corner is the Final Demand rectangle, and let us denote the sum of its elements as F. The one in the lower left corner is the Value-added or Income rectangle, and let us denote the sum of its elements as V. The cornerstone, the fundamental theorem of national accounting is that F = V always and by definition. Before reading the next paragraph, where I will explain why this identity always holds, take a moment to see if you cannot prove it for yourself from what was said in the previous paragraph.

Well, I expect you were successful in your own proof, but to be sure no one gets confused here, let me add my own version. Let us denote by Z the sum of all the intermediate sales from one industry to another, that is, the sum of all the elements in the upper left square where the industry rows cross the industry columns. (In the example, Z = 1180.) Now, from the fact the sum of each industry row is equal to the sum of the corresponding industry column – by the definition of profits – the sum of all industry rows is the sum of all industry columns, that is, Z + F = Z + V. Now, just subtracting Z from both sides gives F = V. We shall refer to this equality as the "product = income identity."

The first input-output tables were made by Leontief and published in 1937 and, more completely, in 1939. These tables were not, however, integrated into the NIPA, for that was not the main direction of Leontief's thinking. That integration was accomplished in England during World War II. It is said that Churchill complained to Keynes that all the economists were producing was theories when he needed to know how big the economy was in order to plan the war effort. In other words, the same question that Petty had dealt with three centuries earlier – and that had plagued Nicholas – was back. Keynes recommended that a young scholar by the name of Richard Stone be put to work on the question. Stone produced not only accounts but pushed further the integration of the input-output table with the accounts. After the war, he produced for the United Nations a volume on the concepts and methods of making national accounts. This work defined the "System of National Accounts" (SNA) which has now spread over most of the world, for no country can afford the luxury of remaining ignorant of what is going on in its economy.

During World War II, the embryonic accounts in the United States had proven very useful, and in 1947 a significant expansion of the accounts appeared, now on a quarterly basis. More and more detail has been gradually added to the NIPA, and they have become central to business forecasting as well as to economic policy. Currently, about 1000 series are available quarterly and about 5600 annually. It is pleasant but hardly surprising to note that Kuznets, Leontief, and Stone all received Nobel prizes, separately, for their work. The United States, however, having come out with its own system of accounts before the appearance of Stone's volume, has not adopted the SNA. There is a program to rework the accounts into the SNA format. The tables presented here, however, are the present U.S. standard, which, it must be said, are easier to follow but less elegant than the SNA.

The "product = income" identity is expressed in Account 1 of the NIPA shown in the following pages. The top of Account 1 shows the components of V, corresponding to the row sums of the income rows; and the bottom half shows the composition of F, corresponding to the column sums of the final demand columns. Obviously, a magnitude as important as F or V is going to have a name, and the name is Gross domestic product (GDP). Account 1, therefore, shows that Gross domestic product is the same, whether computed as the sum of incomes or as the sum of the -- very different -- final demands. Well, to be more precise, the account shows that the NIPA statisticians can't quite make the numbers from the two entirely different sets of sources match. In the upper half of the account, there is the small entry "Statistical discrepancy." It must be added to the income sources to get the final demand (or product) totals.

The product = income identity has further consequences. Let us write it as

$$\text{PersCons} + \text{Inv} + \text{Exp} - \text{Imp} + \text{Gov} = \text{ConCap} + \text{Lab} + \text{Prof} + \text{NetInt} + \text{IndTax}.$$

By simply re-arranging the terms, it implies

$$\text{Inv} + (\text{Exp} - \text{Imp}) = \text{ConCap} + \text{Lab} + \text{Prof} + \text{NetInt} + \text{IndTax} - \text{PersCons} - \text{Gov}$$

The first term on the left is domestic investment, the second is net exports, and the entire right side represents national saving, for it is the sum of all sources of income less the consumption of persons and governments. Thus, we have the rather surprising result that if we know how much our saving is and how much our investment is, we can deduce by how much our exports exceed or fall short of our imports. Since the quantity Exp - Imp represents an increase in national wealth abroad, it is also known as net foreign investment. It appears near the bottom of Account 4, where it is defined as the residual required to make the bottom of the account (imports) equal to the top of the account (exports). Now note in Account 5, the Saving = Investment account, that exactly the same item is added to Gross private domestic investment to equal the sum of all sources of saving.

This Account 5 is essentially the above identity after some algebraic transformation of the right hand side. To see what has been done, let us first use the equation

```
       Prof = ProfitTax + Dividends + UndistribProfits
```

to rewrite the right-hand side of the above identity as

```
       ConCap + Lab + ProfitTax + Dividends + UndistribProfits + NetInt + IndTax -
PersCons - Gov
```

and then rewrite the whole expression as

```
       + Lab + Dividends + GovTr + (NetInt + IntPBP + IntPBG - IntRBG) - PersTax -
Contrib - PersCons - IntPBP
       + ConCap + UndistribProfits
       + IndTax + PersTax + ProfitTax + IntRBG + Contrib - GovTr - Gov - IntPBG.

where
       GovTr         = Government transfer payments to persons
       IntPBP               = Interest paid by persons
       IntPBG               = Interest paid by government
       IntRBG               = Interest received by government
       Contrib              = Contributions for social insurance
```

Note that if a new term was added at one place, it was subtracted at another, so that the total expression has the same value. In the first line, the expression in parenthesis is, as seen in the bottom half of Account 2, is Personal interest income. ("Net interest" is interest paid by business less interest received by business; this term is therefore interest paid by business, government and persons minus interest received by business and government. Since all interest paid must be received by someone, what is left must be received by persons.) Thus, the value of the first line is simply Personal income - Personal taxes - Personal outlays = Personal saving. It is worked out in Account 2 and the final result entered in the "saving" half of Account 5. The second line is just Business saving; it has no separate account to work it out, so the two items in the line appear directly in Account 5. The value of the third line is the Government surplus or saving. It is worked out in Accounts 3A and 3B and the results for the two levels of government separately are shown in Account 5.

Working out the conceptual basis of the NIPA was clearly no small accomplishment. The empirical implementation is even more impressive. After the early work by Kuznets and Leontief, this work was taken over by the Department of Commerce where a long succession of unsung but unusually dedicated and competent statisticians have labored to find ever better ways of finding numbers to match the many concepts in the NIPA. A very condensed statement of the sources used ran 18 pages of fine print in the July 1992 issue of the *Survey of Current Business*. In the process of making the NIPA, many details emerge which are not necessary for the identities but are nonetheless valuable. Some are shown in the tables shown here. The lines where there is a +, -, or = sign to the left of the name of the item are the components of the identities. Items without such a sign are simply showing some of the supporting detail in the accounts. There is much further detail, especially in the annual accounts.

Account 1. National Income and Product Account

Gross Domestic Product by Type of Income

	1985	1990	1995	1997
	----	----	----	----
+ Compensation of employees	2425.7	3352.8	4215.4	4703.5
Wages and salaries	1995.7	2757.6	3442.6	3878.5
Government	373.5	517.2	623.0	665.3
Other	1622.2	2240.3	2819.6	3213.2
Supplements to wages and salaries	430.0	595.2	772.9	825.0
Employer contributions for social insurance	226.9	294.6	366.0	408.4
Other labor income	203.1	300.6	406.8	416.6
+ Proprietors' income with inventory valuation and capital consumption adjustments	268.6	374.0	489.0	544.5
+ Rental income of persons with capital consumption adjustment	48.1	61.0	132.9	148.0
+ Corporate profits with inventory valuation and capital consumption adjustments	303.9	397.1	649.9	807.4
Corporate profits with inventory val. adjustment	230.5	358.2	598.4	548.7
Profits before tax	229.9	371.7	622.6	545.4
Profits tax liability	96.5	140.5	213.2	186.0
Profits after tax	133.4	231.2	409.4	359.4
Dividends	92.8	151.9	264.4	336.1
Undistributed profits	40.6	79.4	145.1	109.7
Inventory valuation adjustment	0.5	-13.5	-24.2	5.6
Capital consumption adjustment	73.5	38.9	51.6	69.7
+ Net interest	337.2	467.3	425.1	448.1

```
  = National income                                      3383.4
4652.1  5912.3  6651.4

  + Business transfer payments                             20.9
26.6    32.2    35.4
  + Indirect business tax and nontax liability            329.6
442.6   582.8   619.4
  - Subsidies less current surplus of gov't enterprises    21.9
25.3    25.3    26.1
  + Consumption of fixed capital                          619.6
696.0   739.3   772.1
  = Gross national income (4_11)                         4198.7
5747.5  7298.9  6069.4

  + Statistical discrepancy                                 2.4
17.4   -28.2   -86.0
  = Gross national product                               4201.0
5764.9  7270.6  8061.8

  - Receipts of factor income from rest of the world      108.1
177.5   222.8   262.1
  + Payments of factor income to rest of the world         87.7
156.4   217.6   281.2
  = GROSS DOMESTIC PRODUCT                               4180.7
5743.8  7265.4  8081.0

=================================================================
=======================
 + Personal consumption expenditures                     2704.8
3839.3  4957.7  5488.1
     Durable goods                                        361.0
476.5   608.5   659.1
     Nondurable goods                                     927.6
1245.3  1475.8  1592.1
     Services                                            1416.1
2117.6  2873.4  3236.9
 + Gross private domestic investment                      715.1
799.7  1038.2  1240.9
     Fixed investment                                     688.9
791.7  1008.1  1172.6
       Nonresidential                                     502.0
575.9   723.0   845.4
         Structures                                       193.3
200.8   200.6   229.9
         Producers' durable equipment                     308.6
375.1   522.4   615.5
       Residential                                        187.0
215.7   285.1   327.3
```

```
     Change in business inventories                        26.2
 8.0    30.1    68.3
 + Net exports of goods and services                     -114.2
-71.3   -86.0  -100.7
     Exports                                              303.0
557.3   818.4   958.0
     Imports                                              417.3
628.6   904.4  1058.8
 + Government purchases                                   875.0
1176.2  1355.5  1452.7
     Federal                                              410.1
503.6   509.6   523.8
       National defense                                   312.4
373.1   344.6   350.4
       Nondefense                                          97.7
130.4   165.0   173.4
     State and local                                      464.9
672.5   846.0   929.0
 = GROSS DOMESTIC PRODUCT                                4180.7
5743.8  7265.4  8081.0
=================================================================
=============================
```

Account 2. Personal Income and Outlay Account

	1985	1990	1995	1997
	----	----	----	----
+ Personal tax and nontax payments	437.6	624.8	795.1	988.7
+ Personal outlays	2795.8	3958.1	5101.1	5660.8
Personal consumption expenditures	2704.8	3839.3	4957.7	5488.1
Durable goods	361.0	476.5	608.5	659.1
Motor vehicles and parts	175.7	210.4	254.8	263.3
Furniture and household equipment	126.3	176.0	240.1	267.4
Other	59.0	90.1	113.6	128.5
Nondurable goods	927.6	1245.3	1475.8	1592.1
Food	466.5	630.5	735.1	776.5
Clothing and shoes	152.1	205.9	254.7	277.2
Gasoline and oil	97.2	109.2	114.4	124.6
Fuel oil and coal	13.5	12.0	10.2	10.8
Other	198.2	287.6	361.4	402.9
Services	1416.1	2117.6	2873.4	3236.9
Housing	407.0	586.3	750.3	826.5
Household operation	180.3	226.3	300.7	328.6
Electricity and gas	88.8	98.7	119.5	127.0
Other household operation	91.4	127.6	181.1	201.6
Transportation	100.0	143.7	203.1	236.3
Medical care	321.8	537.7	772.8	855.1
Other	407.0	623.5	846.5	990.5
Interest paid by persons	83.2			

```
108.8   128.6   154.8
     Personal transfer payments to rest of the world (n    7.8
 9.9    14.8    17.8
 + Personal saving                                        216.4
221.3   254.6   224.7
 = PERSONAL TAXES, OUTLAYS, AND SAVING                   3449.8
4804.2  6150.8  6874.2
=================================================================
============================

 + Wage and salary disbursements                         1995.9
2757.5  3429.4  3877.3
     Commodity-producing industries                       620.7
754.2   864.3   960.2
       Manufacturing                                      468.9
561.2   648.4   705.9
     Distributive industries                              476.5
634.1   783.1   876.2
     Service industries                                   525.0
852.1  1159.0  1375.5
     Government                                           373.8
517.2   623.0   665.3
 + Other labor income                                     203.1
300.6   406.8   416.6
 + Proprietors' income with inventory valuation           268.6
374.0   489.0   544.5
    and capital consumption adjustments (CCAdj)
 + Rental income of persons with CCAdj                     48.1
61.0   132.9   148.0
 + Personal dividend income                                88.3
142.9   251.9   321.5
 + Personal interest income                               508.4
704.4   718.9   768.8
     Net interest                                         337.2
467.3   425.1   448.1
     Interest paid by persons                              83.2
108.8   128.6   154.8
     Interest paid by government                          195.9
268.6   314.1   319.1
     -Interest received by government                     107.9
140.4   148.9   154.1
     Interest paid by persons                              83.2
108.8   128.6   154.8
 + Transfer payments to persons                           486.6
687.8  1015.0  1121.2
     Old-age, survivors, disability, and                  253.4
352.0   507.8   566.7
       health insurance benefits
     Government unemployment insurance benefits            15.7
18.1    21.5    21.8
     Veterans benefits                                     16.7
17.8    20.8    22.4
     Government employees retirement benefits              66.6
94.5   133.6   153.4
     Other transfer payments                              134.1
205.3   331.3   356.9
       Aid to families with dependent children             15.4
19.8    23.3    18.7
       Other                                              118.7
185.5   308.0   338.2
 - Personal contributions for social insurance            149.0
223.9   293.1   323.7
 = PERSONAL INCOME                                       3449.8
4804.2  6150.8  6874.2
=================================================================
===========================
```

Account 3A. Federal Government Receipts and Expenditures Account

	1985	1990	1995	1997
+ Purchases	410.1	503.6	509.6	523.8
National defense	312.4	373.1	344.6	350.4
Nondefense	97.7	130.4	165.0	173.4
+ Transfer payments (net)	379.0	513.3	720.9	795.5
To persons	366.9	500.0	709.4	782.2
To rest of the world (net)	12.2	13.3	11.5	13.3
+ Grants-in-aid to State and local governments	100.3	132.4	211.9	224.2
+ Net interest paid	126.9	179.9	224.8	230.2
Interest paid	153.8	208.2	250.0	254.5
To persons and business	130.7	167.1	188.7	123.3
To rest of the world (net)	23.1	41.0	61.3	67.4
- Interest received by government	26.9	28.3	25.2	24.3
+ Subsidies less current surplus of gov't enterprises	25.2	32.4	36.4	38.3
Subsidies	22.8	28.1	33.7	34.2
- Current surplus of government enterprises	-2.4	-4.3	-2.8	-4.2
- Wage accruals less disbursements	-0.2	0.1	0.0	0.0
= Federal Expenditures	974.2	1284.5	1637.6	1752.2
+ Surplus or deficit (-), NIPA basis	-162.9	-154.6	-174.4	-25.8
In social insurance funds	33.1	80.1	54.1	63.7
Other	-196.0	-234.8	-228.6	-71.7
= FEDERAL GOVERNMENT EXPENDITURE AND SURPLUS	811.2			

```
1129.8  1463.2  1281.6
=================================================================
===========================

 + Personal tax and nontax receipts                       343.7
485.7   605.8   774.4
     Income taxes                                         336.7
472.7   588.7   751.1
     Estate and gift taxes                                  6.4
11.6    14.9    20.6
     Nontaxes                                               0.6
 1.4     2.2     2.7
 + Corporate profits tax accruals                          76.2
117.9   182.1   212.8
     Federal Reserve banks                                 17.8
23.6    23.4    15.9
     Other                                                 58.5
94.4   158.7   142.1
 + Indirect business tax and nontax accruals               58.2
65.1    93.5    91.3
     Excise taxes                                          34.5
35.1    58.1    58.7
     Customs duties                                        12.2
17.5    19.4    19.8
     Nontaxes                                              11.5
12.4    16.1    12.9
 + Contributions for social insurance                     333.1
461.1   581.8   645.9
 = FEDERAL GOVERNMENT RECEIPTS                            811.2
1129.8  1463.2  1281.6
=================================================================
============================
```

Account 3B. State and Local Receipts and Expenditures Account

	1985	1990	1995	1997
	----	----	----	----
+ Purchases	464.9	672.5	846.0	929.0
Consumption	382.7	550.1	698.6	762.9
Investment	82.3	122.5	147.4	166.1
+ Transfer payments to persons	101.8	166.5	280.6	311.8
+ Net interest paid	-38.9	-51.7	-59.6	-65.2
Interest paid	42.1	60.4	64.1	64.6
Less: Interest received by government	80.9	112.0	123.7	129.8
- Dividends received by government\1\	4.5	9.0	12.5	14.6
+ Subsidies less current surplus of gov't enterprises	-3.3	-7.1	-11.2	-12.2
Subsidies	0.3	0.4	0.3	0.3
- Current surplus of government enterprises	3.6	7.5	11.5	12.6
- Wage accruals less disbursements	0.0	0.0	0.0	0.0
= Expenditures	437.8	648.8	895.9	982.6
+ Surplus or deficit (-), NIPA basis	91.0	80.1	103.0	80.2
Social insurance funds	47.0	59.9	70.5	71.4
Other	44.0	20.2	32.5	26.7
= STATE AND LOCAL GOVERNMENT EXPENDITURES AND SURPLUS	528.7	729.0	999.0	812.5

	1985	1990	1995	1997
+ Personal tax and nontax receipts	93.9	139.1	189.4	214.3
Income taxes	72.1			

```
106.3   140.3   159.8
      Nontaxes                                                9.9
15.5    26.6    31.0
      Other                                                  11.9
17.2    22.4    23.5
 + Corporate profits tax accruals                            20.2
22.5    31.1    39.5
 + Indirect business tax and nontax accruals                271.4
377.6   489.3   528.1
     Sales taxes                                            131.1
183.2   239.4   257.4
     Property taxes                                         107.0
155.4   197.4   208.7
     Other                                                   33.3
38.9    52.5    62.0
 + Contributions for social insurance                        42.8
57.4    77.2    86.2
 + Federal grants-in-aid                                    100.3
132.4   211.9   224.2
 = STATE AND LOCAL GOVERNMENT RECEIPTS                      528.7
729.0   999.0   812.5
=================================================================
=============================
```

Account 4. Foreign Transactions Account

	1985	1990	1995	1997
+ Exports of goods and services	303.0	557.3	818.4	958.0
Exports of merchandise	222.3	398.5	583.9	686.5
Foods, feeds, and beverages	24.5	35.2	50.5	51.1
Industrial supplies and materials	59.4	101.8	141.3	152.7
Durable goods	16.9	35.7	49.8	55.1
Nondurable goods	42.5	66.1	91.4	97.7
Capital goods, except automotive	79.3	152.6	233.8	294.5
Civilian aircraft, engines, and parts	13.6	32.2	26.1	41.4
Computers, peripherals, and parts	14.7	25.9	39.7	49.5
Other	51.1	94.4	168.0	203.6
Automotive vehicles, engines, and parts	25.0	36.5	61.8	73.6
Consumer goods, except automotive	14.6	43.7	64.4	77.5
Durable goods	6.4	23.8	32.8	40.0
Nondurable goods	8.2	20.0	31.6	37.6
Other	19.4	28.9	32.1	37.1
Durable goods	9.7	14.4	16.1	18.5
Nondurable goods	9.7	14.4	16.1	18.5
+ Receipts of factor income from the rest of the world	108.1	177.5	222.8	262.1
+ Capital grants received by the United States (net)	0.0	0.0	0.0	0.0
= RECEIPTS FROM THE REST OF THE WORLD	411.1	734.8	1041.2	1220.2

```
==============================

 + Imports of goods and services                              417.3
628.6   904.4  1058.8
     Imports of merchandise                                   343.3
508.0   757.6   888.8
       Foods, feeds, and beverages                             21.9
26.4    33.2    39.7
       Industrial supplies and materials                       59.2
78.1   119.9   135.1
         Durable goods                                         30.5
38.4    59.8    69.2
         Nondurable goods                                      28.7
39.6    60.1    65.9
       Petroleum and products                                  51.4
62.3    56.1    72.1
       Capital goods, except automotive                        61.3
116.1   221.4   254.1
         Civilian aircraft, engines, and parts                  5.3
10.5    10.7    16.6
         Computers, peripherals, and parts                      8.4
22.9    56.3    70.1
         Other                                                 47.6
82.7   154.4   167.3
       Automotive vehicles, engines, and parts                 64.9
88.5   123.8   141.3
       Consumer goods, except automotive                       66.3
105.0   159.9   192.9
         Durable goods                                         38.3
55.7    83.7    98.4
         Nondurable goods                                      28.0
49.3    76.2    94.5
       Other                                                   18.2
31.7    43.2    53.5
         Durable goods                                          9.1
15.8    21.6    26.8
         Nondurable goods                                       9.1
15.8    21.6    26.8

 + Payments of factor income to the rest of the world          87.7
156.4   217.6   281.2

 + Transfer payments to rest of the world (net)                23.1
28.4    33.6    39.4
     From persons (net)                                         7.8
 9.9    14.8    17.8
     From government (net)                                     12.2
13.3    11.5    13.3
```

```
     From business                                          3.1
 5.2     7.3     8.3

 + Net foreign investment                                -117.0
-78.7  -114.3  -115.8

 = PAYMENTS TO THE REST OF THE WORLD                      411.1
734.7  1041.2  1263.6
=================================================================
=============================
```

Account 5. Gross Savings and Investment Account

```
                                                           1985
1990    1995    1997
                                                           ----
----    ----    ----
 + Gross private domestic investment                      715.1
799.7  1038.2  1240.9
 + Net foreign investment                                -117.0
-78.7  -114.3  -115.8
 = GROSS INVESTMENT                                       598.1
721.0   923.8  1125.1
=================================================================
=============================
 + Personal saving                                        216.4
221.3   254.6   224.7
 + Wage accruals less disbursements                        -0.2
 0.1    13.1     1.2
 + Undistributed corporate profits with IVA and CCAdj     -33.4
54.0   117.7    34.3
 + Consumption of fixed capital                           486.6
651.5   796.8   867.9
 + Federal surplus                                       -162.9
-154.6  -174.4   -25.8
 + State and local surplus                                 91.0
80.1   103.0    80.2
 + Capital grants received by the United States (net)       0.0
 0.0     0.0     0.0
 + Statistical discrepancy                                  2.4
17.4   -28.2   -86.0
 = GROSS SAVING AND STATISTICAL DISCREPANCY               598.1
721.0   923.8  1125.1
=================================================================
=============================
```

2. Price indexes

The account shown above are all in current prices, that is, the data for 1980 are in 1980 prices, while those for 1990 are in 1990 prices, and so on. The accounts in current prices are fine for showing the relations among various items in the same year. But they are of limited usefulness for comparing variables across years, because inflation has moved up many prices. How to measure price changes and put the accounts into "constant" prices is one of the most vexing subjects of applied economic statistics.

xxx more xxx