

CHAPTER 2

USEFUL TOOLS FOR PRACTICAL BUSINESS FORECASTING



INTRODUCTION

Statistical theories may provide the necessary base for estimating regression equations, but any modeler immediately faces the following practical questions:

- Where do I find the data?
- How should the data be collected and organized?
- What is the best method for presenting the data graphically?
- What software should be chosen for calculating the regressions and building the model?

There are hundreds if not thousands of data sources, and many graphical packages and econometric software that will perform well. However, this author and his colleagues have found that model builders are not helped by being presented with an encyclopedic list of possible sources. Instead, this text uses an eclectic approach, offering the data sources and software packages that have proven most useful in our own work.

Today, much of the relevant data can be obtained from the Internet. This book provides addresses with the following caveat: they change frequently, and sometimes updated links are not given. Nonetheless, by briefly describing the relevant data sources, it should be straightforward to track them down even if the addresses have changed.

Before using the data for any type of analysis, it is always best to check for errors. Few things are more annoying than spending a great deal of time building and testing a model, only to find the underlying data are flawed. In most cases, the data will be identified as being seasonally adjusted or unadjusted, but different methods of seasonal adjustment could cause a variety of problems when the model is being used for forecasting. In addition, it is always prudent to check for outlying observations, and for missing or incomplete data, which could seriously distort the results of the model.

The software package used to estimate the models given in this text is EViews, a comprehensive program that is useful both for estimating regressions and building models. For those who plan to engage in large-scale data collection and model building, a program with the capacity and power of EViews is essential. On the other hand, many model builders develop equations with relatively few observations, and would prefer to link their models to spreadsheet analysis that has already been developed in Excel. While Excel is not recommended for heavy number-crunching, it can be a very useful tool for small models. However, the examples given in the text are based on EViews.

2.1 TYPES AND SOURCES OF DATA

The data that economic model builders use to generate forecasts can be divided into three principal categories: time-series, cross-section, and panel data. Most forecasting models use time-series data. A time series is a sequence of data at equidistant intervals where each point represents a certain time period (e.g., monthly, quarterly, or annually). Examples include quarterly data for consumption, monthly data for industrial production or housing starts, daily data for the stock market, annual data for capital spending, quarterly data for individual company sales and profits, or monthly levels of production and inventories.

Most econometric and forecasting books cover “regression models” and “time-series models.” The first category includes the construction of models based on underlying economic theory; which are generally known as structural models. The second category incorporates models that relate the data to its previous values, time trends, seasonal adjustment factors, and other exogenous variables. Since no attempt is made to provide an underlying theory, these are known as non-structural models. As is shown later, superior forecasts are often generated by combining these two methods.

2.1.1 TIME-SERIES, CROSS-SECTION, AND PANEL DATA

Admittedly, use of the term “time series” to describe two different phenomena can sometimes be confusing. Time-series *data* are used in both regression models and time-series *models*. Time-series *data* refer to a time sequence of events regardless of the type of model in which they are used. Most of the material in this book will utilize time-series data. Part II of the text covers regression models, while Part III discusses time-series models; both are based on time-series data.

Cross-section data represent a snapshot of many different observations taken at a given time. The decennial census of population data are often used for cross-section analysis; for any given census year, statistical relationships can be

used to estimate the correlation between income and education, location, size of family, race, age, sex, and a whole host of other variables.

Internal Revenue Service data are often used by the Congressional Budget Office to determine how various changes in the tax laws would affect individuals at various levels of income distribution; e.g., whether a particular tax cut would mainly benefit the “rich” or the “poor.” Consumer survey data reveal the proportion of income that is spent on various goods and services at different levels of income. For example, economists might want to examine the behavior of a group of consumers to determine the level of their income, saving, and pattern of consumption (the relative amounts spent on food, rent, cars, vacations, etc.) at some specific time, say June 1995. Similar surveys can be used to determine the mix of goods purchased by, say, consumers in New York City compared to Denver, Colorado. At a more detailed level, individual companies use cross-section analysis to help determine who buys their products at department stores and supermarkets. Data on personal health collected at a specific time can be used to reveal what type of individual has the greatest risk for various diseases based on age, income, eating habits, use of tobacco and alcohol, parental health history, and other factors.

Much econometric work is based on cross-section data. For example, researchers might be interested in finding out how different types of consumers reacted to a tax change in the past. Economists have used cross-section data to determine whether the overall growth rate in a given country is due to government policies, the saving and investment ratio, education of the population, and many other factors. Financial advisors might be interested in determining the probability that a municipal bond issue would default, based on per capita income of the issuing municipality, age/sex/race characteristics, projects for which the money will be used, existing tax base and growth in that base, and so on. There are many more useful examples of how cross-section data can be used to predict various events, some of which will be used as examples later in this book.

Panel data refers to the reexamination of cross-section data with the same sample at different periods of time. For example, the problem with the June 1995 data might be that individuals buy a new car on average only once every four years (say), so that month might not have been typical. Thus the same people could be asked about their income, saving, and consumption in January 1997, and at other periods. Over a longer period of time, the spending patterns of these individuals could be tracked to help determine how much is saved at different levels of income, whether upper-income people spend a larger proportion of their income on housing, transportation, or medical care, or a host of other items. Panel data could also be used to determine whether individuals who started smoking cigarettes at a young age continued to smoke throughout their lives, whereas those who started smoking later found it easier to quit. These data could also help determine whether an increase in excise taxes on cigarettes has a greater effect in reducing smoking in the long run than in the short run.

2.1.2 BASIC SOURCES OF US GOVERNMENT DATA

Those who build a forecasting model using time-series data generally use government data even if they are predicting individual industry or company sales. Unless these forecasts are entirely driven by technology, they will depend on the level of economic activity both in the US and abroad.

The main US government data search engine (see section 2.2) lists 70 agencies that supply US economic data. However, for most economic forecasting needs, the main data sources are the Bureau of Economic Analysis (BEA), the Bureau of the Census, the Bureau of Labor Statistics (BLS), and the Board of Governors of the Federal Reserve System (Fed). Other important government sources of data include the Statistics of Income Division of the Internal Revenue Service; the Economic Research Service of the Department of Agriculture, and the Energy Information Administration of the Department of Energy. Since this is a brief book on forecasting rather than the sources of government data, the discussion at this point will be limited to the first four agencies.

The National Income and Product Accounts (NIPA) are prepared by BEA, which is part of the Commerce Department. The figures for current dollar and inflation-adjusted GDP,¹ consumption and investment, and personal and corporate income are all calculated and reported by BEA. In addition, BEA offers comprehensive data on state and county personal income and employment by detailed industry classification.

BEA processes and compiles data that are collected by various other government agencies. Most of the series that serve as inputs for NIPA are collected by the Bureau of the Census, which is also part of the Commerce Department. The Census Bureau is perhaps best known for its decennial count of all people in the country, but that is only a small part of its total activity. Most of the monthly reports on economic activity issued by the government are published by Census. These reports include data for manufacturers shipments, orders, and inventories; wholesale and retail sales and inventories; housing starts and construction put in place; and exports and imports. While most of the NIPA figures (except consumption and income) are quarterly data, data in the census publications listed here are all available on a monthly basis. Census also publishes the *Quarterly Report for Manufacturing Corporations*, which provides data for all major income statement and balance sheet items for all major manufacturing industries by asset size.

¹ Before 1996, figures were available in current and constant dollars. However, the methodological revisions introduced by BEA in 1996 switched to the use of chain-weighted figures to adjust for inflation, which essentially differ from constant dollars in that the weights are reset each year. The practical impact of this change is that the components of aggregate demand with decreasing prices, notably computer purchases, rise less rapidly than the constant-dollar figures, so the distorting influence on total aggregate demand is smaller.

Data for wages, prices, employment, unemployment, productivity, and labor costs are issued monthly by the BLS, which is part of the Labor Department. The BLS data have the biggest short-term impact on financial markets. The Employment and Earnings Report, which contains data on employment, unemployment, and wage rates; and the producer price index and consumer price index (PPI and CPI) are the most closely watched economic indicators by financial markets. The BLS also compiles monthly data on state and metropolitan area employment and unemployment.

The fourth major source of government data is the Fed. As would be expected, most of its reports cover monetary variables, including the money supply, bank assets and liabilities, interest rates, and foreign exchange rates. However, the Fed also provides figures for industrial production and capacity utilization for the overall economy and by detailed manufacturing industry.

Most of the key numbers that economists use are collected and issued in a monthly release called, appropriately enough, *Economic Indicators*, which is issued by the Council of Economic Advisers. It contains slightly over 500 series of economic data and can be purchased from the Government Printing Office for \$55.00 per year (as of 2002). Updated data are also available on the Internet at www.access.gpo.gov/congress.cong002.html.

Economic Indicators is designed to present the most recent data, so it does not contain very much historical data. That can be found in the annual issues of the *Economic Report of the President*, a useful source for annual government data, although monthly and quarterly data are presented only for recent years.

The *Survey of Current Business*, published by the Commerce Department, contains comprehensive NIPA tables and a few other series, but it contains far less data since budget cuts stripped thousands of series from its tables. All the data in that publication can be found by accessing the BEA home page at www.bea.doc.gov and following the directions. Data for GDP by industry, state personal income, and a variety of regional economic data can be obtained by purchasing CD-ROMs at prices ranging from \$20.00 to \$35.00, and may be purchased from the BEA order desk. This website also allows viewers to access individual tables in the NIPA accounts and data for individual US states.

Those who want to obtain the government data immediately can subscribe to a Commerce Department service known as STAT-USA, which provides all key economic reports within minutes of their release time by the particular government agency. It's not free: having the data available immediately means subscribing for \$175 or more per year. It is well worth it for those who follow the data closely and depend on those numbers for making financial market decisions, but for those who are just building a quarterly or annual model, time is generally not of the essence.

For those not familiar with the scope of government data, the best place to start for "one-stop" shopping is the *Statistical Abstract of the United States*, which is issued annually by Census. It contains approximately 1500 tables of data, most of it pertaining to the US economy, and also lists Internet addresses for

the 33 major sources of Federal government data. The numbers found in the *Statistical Abstract* are collected from over 200 sources. Census sells a CD-ROM for each annual edition; the major drawback is that most of the series are given for only a few years, so to collect historical time series one must go to the original source or look through back issues of the *Statistical Abstract*. Census also sells CD-ROMs with economic data series for states, metropolitan areas, and individual counties.

2.1.3 MAJOR SOURCES OF INTERNATIONAL GOVERNMENT DATA

So far we have looked at only US government data, whereas many business applications increasingly rely on foreign data. For those who are building a specific model of a foreign country, most other industrialized countries have central sources of data similar to the US. In most cases, however, model builders will want general summary statistics for a wide variety of countries, which might be used (for example) in determining which countries represent the fastest growth potential, the best place for locating new plants, or the biggest risk factor in terms of depreciating currencies.

There are two general sources for international data. The first is the Organization for Economic Cooperation and Development (OECD), which is headquartered in Paris but has an office with most of their statistical documents in Washington, DC. It publishes several volumes of data, mainly NIPA figures and basic data for employment, production, and prices. Most of the data series are monthly or quarterly, but they are available only for the 29 OECD countries.

The International Monetary Fund (IMF), also in Washington, DC, provides data for over 170 countries, but almost all of the series are annual. As might be expected, its series concentrate on monetary data, balance of payments figures, and exchange rates, with relatively little data for real output, production, prices, and employment.

Specific data for Canada can be obtained from Statistics Canada, at www.statcan.ca/. Data for Europe issued by the European Central Bank can be found at www.ecb.int/. Data for Japan are available from the Bank of Japan at www.boj.or.jp/en/. Eurostat has a website with a wide variety of economic indicators for countries that have joined together in the euro; that information can be found at www.europa.eu.int/comm/eurostat.

For those who like to keep up to date on international data at a relatively modest cost, *The Economist* magazine carries key economic series for major countries in each weekly issue, which can be accessed at www.economist.com. Some data are available to all users; most are restricted to those who subscribe to the print version of that publication.

The Census Bureau has a comprehensive database for 227 countries and areas of the world for those interested in demographic and socioeconomic

data: population, birth and death rates, literacy, and so on; it also contains statistics for labor force, employment, and income. This can be found at www.census.gov/ipc/www/idbnew.html.

If you are looking for specific economic data that are not found at the above sources, the Dallas Fed has a comprehensive set of links to international data. This can be accessed at www.dallasfed.org/htm/data/interdata.html. One of the most useful links will take you to Statistical Data Locators, a comprehensive list of organizations that is compiled by NTU Library at Singapore.

The Office of Productivity and Technology at BLS also publishes monthly data for the CPI and unemployment rates for major countries, plus figures for unit labor costs in manufacturing and per capita GDP for most OECD countries. As is the case for other BLS series, these can be accessed at www.bls.gov.

Those figures cover recent years; for older data, the standard source is by Robert Summers and Alan Heston, entitled *The Penn World Table Mark 5: An Expanded Set of International Comparisons, 1950–88*. This article originally appeared in the *Quarterly Journal of Economics* in 1991, but the data can also be obtained from the National Bureau of Economic Research (NBER) in New York. Actual data can be downloaded at <http://datacentre.chass.utoronto.ca:5680/pwt/index.html>. While these data are very comprehensive and are used for many international research studies, most of the series are only available up to 1994 and are not updated very frequently. As of 2001, the latest available version, 5.6, was released in January, 1995.

2.1.4 PRINCIPAL SOURCES OF KEY PRIVATE SECTOR DATA

While there are myriad sources of private sector data, many of them are either available only to members of specific organizations, or are sold at a very high price. This book does not offer a survey of these private sector databases; comments here are restricted to data that are generally available at a zero or modest price. These sources can be divided into the following categories:

- financial market data
- individual company data
- consumer behavior
- housing surveys
- manufacturing sector surveys
- individual industry data.

Except for financial data, the first place to look is often the *Statistical Abstract*, which has recent figures for most series and provides the source for comprehensive historical data. Although most of their data comes from government sources, about 10 percent of the 1,500 tables, each containing several series, are from private sector sources.

FINANCIAL MARKET DATA

The standard source is the Center for Research on Security Prices at the University of Chicago. However, that huge database is likely to be more than is needed by those who are planning to analyze only a few companies, or need data for only a relatively short period of time. Worden Brothers, which can be accessed at www.TC2000.com, will send a CD-ROM with daily stock market data for up to 15 years at no cost. They hope users will update the data at \$1.00/day, but even for those who do not choose that option, the CD-ROM will supply a great deal of historical data on individual stocks.

INDIVIDUAL COMPANY DATA

The best bet is to access the Web. Hoover's On-Line is one convenient source that has over 50 databases that offer individual company data. One of those databases is Public Register's Annual Report Service, which provides free annual reports for over 3,600 firms. The Securities and Exchange Commission EDGAR file contains all reports that must be filed by public companies; that would be more than most people need, but it can be a valuable resource.

CONSUMER BEHAVIOR

The two key surveys are undertaken by the Conference Board and University of Michigan. The Conference Board is willing to have their data disseminated, and makes much of it available for free or a modest fee. The University of Michigan, on the other hand, is concerned that if they give out their survey results, hardly anyone will pay to subscribe. Nonetheless, all the wire services carry their reports a few minutes after they are released, so the data can be obtained second-hand from various sources. However, the Conference Board is much more customer-friendly. In empirical testing, this author has found relatively little difference between the two series.

HOUSING SURVEYS

The main surveys are undertaken by the National Association of Home Builders and the National Association of Realtors. These surveys contain data for number of homes sold and average price by state and detailed metropolitan area, characteristics of new homes being built, and attitude surveys about the likelihood of consumer purchases in the near term. In both cases, the overall numbers are available for free, while data in the detailed reports can be purchased.

MANUFACTURING SECTOR SURVEYS

The best-known survey is published by the National Association of Purchasing Managers. It is released monthly, based on questionnaires filled out by

approximately 250 purchasing managers about shipments, production, employment, delivery times, and especially prices paid and received. Several regional purchasing managers' indexes are also published, notably for Chicago and New York, but the national survey is generally thought to have a higher level of accuracy and is referenced much more frequently.

INDIVIDUAL INDUSTRY DATA

Some of the major associations that will make their summary data available free or at modest cost include the American Iron & Steel Institute, Association for Manufacturing Technology (formerly Association of Machine Tool Builders), American Petroleum Institute, Electronics Industry Association, Dataquest Gartner (for computer shipments and revenues), and the Semiconductor Industry Association.

2.2 COLLECTING DATA FROM THE INTERNET

Many model builders want to obtain complete historical series of quarterly or monthly data without having to type them in by hand. There are essentially three choices. First, you can pull each series off the Web using the cut and paste routines; the major sources of data from the Internet are discussed in this section. Second, you can order disks or CDs from each of the government agencies. Third, you can pay someone else to do the heavy lifting by purchasing a comprehensive database from some commercial vendor. The databases used in conjunction with EViews are compiled by Haver Analytics. Other commercial vendors offer similar databases, but at somewhat higher prices. Unless otherwise stated, the data referenced in this text were either collected by the author directly or are found in the Haver Analytics database. The basic Haver database covers only US data except for a few foreign interest and exchange rates. Comprehensive foreign data can be purchased from OECD or IMF either in printed form or on CD-ROMs.

The section on collecting data from the Internet could be an entire monograph. However, the purpose is not to list all, or even most, of the sources of economic data available on-line. It is to provide a comprehensive but nonetheless compact directory for finding most of the data that are likely to be useful in building econometric models.

For those who know what data series they want, and know the government or private sector source for that data, the obvious choice is to proceed directly to that website. If you don't know who publishes the data, or aren't sure what series you want, several comprehensive data sites on the Web are recommended. The principal sources of US and international public sector data are as follows (website addresses were current as of 2001). The sites that combine many databases are listed in increasing order of generality.

- Bureau of Economic Analysis: National income and product accounts, international transactions, regional income and employment. www.bea.doc.gov
- Bureau of the Census: Monthly data for manufacturers shipments, orders, inventories; wholesale and retail trade and inventories; housing starts and construction put in place; monthly foreign trade statistics. www.census.gov
- Bureau of Labor Statistics: Employment and unemployment; CPI and PPI; wage rates, productivity, and unit labor costs. www.bls.gov
- Board of Governors of the Federal Reserve System: money supply, bank balance sheets, interest rates, foreign exchange rates, industrial production. www.bog.gov
- Internal Revenue Service: income tax data. www.irs.gov
- Organization for European Cooperation and Development (OECD): Most key economic series for OECD countries, many on a monthly or quarterly basis. www.oecd.org
- International Monetary Fund (IMF): Many of the same series as OECD, but for over 170 countries. Most data are on an annual basis, and most of the series are monetary, as opposed to real sector variables or prices. www.imf.org

If you want US government data but do not know who publishes it, try www.Fedstats.gov. That contains a comprehensive list of all data published by 70 government agencies, and the search engine is quite thorough. It is highly recommended for those who want to use government data. The search engine also includes a long list of articles written about subjects related to economic data.

There are many comprehensive sites for economic data on the Internet. If you are looking for strictly economic data, the best site is the St. Louis Federal Reserve Bank database, appropriately named FRED. It can be found at www.stls.frb.org/fred.

For those who want to cast their “net” wider and look for data that encompass both economic and other social sciences, one good choice is the business and economics database at the University of Michigan. The address is www.lib.umich.edu/libhome/Documents.center/stats.html.

Finally, if you are looking for a broader range of economic and business data, the following website lists literally hundreds of individual Web-based databases, although some of the links are out of date. That is found at www.mnsfld.edu/depts/lib/ecostats.html.

2.3 FORECASTING UNDER UNCERTAINTY

Statisticians generally distinguish between two distinct types of forecasting models: those where the underlying probability distribution is known, and those where it isn't. The first type includes such examples as poker hands, chances at the roulette wheel, or the correlation between height and weight. If one were able to perform enough experiments to include the entire population, the results

would be known with certainty. Of course that does not mean the outcome of the next event would be known in advance, only the probability that it would occur. However, if enough experiments were performed, the sample mean and variance would approach the population mean and variance. Even more important, all observations are independent, and the underlying probability distribution does not change. No matter how many times in a row you have won or lost at the roulette wheel, the probability of success on the next spin is independent of what previously happened – assuming that the wheel is not “fixed.”

The other type of forecasting model, which is more relevant to business forecasting, occurs when the underlying probability distribution is not known. We think, for example, that consumers spend more when their income rises, and businesses invest more when the real rate of interest declines. Those are certainly reasonable hypotheses and are buttressed by economic theory. However, consider all the factors we don’t know: *how much* consumption will change when income changes, the time lag, other factors that affect income, the fact that the observations are not independent (most people are creatures of habit), and the fact that we don’t know what income will be in the future. Even more important, the relationship between consumption and income may change for the same individuals depending on the economic environment. They may be more optimistic or more pessimistic; they may have recently moved into a larger home and need more furniture, their children may be approaching college age, and a host of other factors.

Over the past century, a large amount of statistical literature has been devoted to the issue of the “best” methods of estimating empirical relationships. The majority of these articles are related to the method of least squares. However, almost all of the tests and relationships are based on assumptions that do not exist in the typical practical business forecasting environment. The major problems can be briefly summarized as follows:

- The data are not normally distributed.
- The residuals are not all independent (the forecasting error in this period is often closely connected with the error last period).
- The independent variables are supposed to be known at the time of forecast, which is generally not the case.
- The data are sometimes inaccurate and subject to substantial revision.
- Finally, and most important, the underlying data generation function may have shifted during the sample period, or – even more damaging – during the forecast period.

In spite of all these drawbacks, the vast majority of economic forecasting models are estimated using least squares, and the examples given in this book will follow this approach. However, emphasis will be placed on adjusting for the fact that the classical least squares criteria often do not occur. For this

reason I will not offer the usual introductory discussion of the statistics, which can be found in many other suitable textbooks. Two texts this author generally uses for supplementary statistical and econometric material when teaching this course are *Econometric Models and Economic Forecasts*, by Robert S. Pindyck and Daniel L. Rubinfeld, and *Econometric Methods* by Jack Johnston and John DiNardo.² The following chapters will develop as much of the outline of the general linear model as is needed as a framework to explore where the actual results differ. However, before turning to the general linear model, it is best to discuss some of the more common terms that will be used throughout the text. The treatment that follows is non-technical.

2.4 MEAN AND VARIANCE

Suppose the same experiment is performed several times, and we take a weighted average of all the outcomes, where the weights are the probabilities. That weighted average is known as the *expected value*, or *mean* of the distribution, and is usually denoted in statistics by μ . It can be defined as follows:

$$\mu_x = E(X) = p_1X_1 + p_2X_2 + \cdots + p_nX_n = \sum_{i=1}^N p_iX_i \quad (2.1)$$

where the p_i are the probabilities associated with events X_i .

The expected value is closely related to, but not the same as, the *sample mean*, which is the *actual* average value one obtains by performing the experiment a certain number of times. The sample mean is denoted as \bar{X} , where

$$\bar{X} = (1/N) \sum_{i=1}^N X_i. \quad (2.2)$$

As the number of experiments increases, the sample mean always approaches its expected value. That is one of the bases of statistical theory. It is a simple matter to show that $E(\bar{X}) = \mu_x$.

In trying to determine the true underlying value of the parameter with sampling, it is also important to measure the *dispersion* around the mean, and determine whether the sample observations are tightly clustered around the mean or are spread out so that they cover almost the entire range of probabilities. The dispersion around the mean is known as the *variance*, which can be defined as

$$\text{Var}(X) = \sigma_x^2 = \sum p_i [X_i - E(X)]^2 \quad (2.3)$$

² Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th edn (Irwin McGraw-Hill, Boston), 1998. Johnston, Jack, and John DiNardo, *Econometric Methods*, 4th edn (McGraw-Hill, New York), 1997. All page numbers and references are to these editions.

where p_i is the probability of each event X_i occurring, and $E(X)$ is the the expected value of X .

Just as we distinguish between the expected value and the sample mean, we can distinguish between the true variance and its sample estimator. However, whereas the sample mean \bar{X} was an unbiased estimator of the expected value, it turns out that the sample variance $(X - \bar{X})^2$ is *not* an unbiased estimator of the variance. Instead, it must be adjusted for what is known as *degrees of freedom*, which equals the number of observations minus the number of variables in the equation. As a result, an unbiased estimate of the variance of a random variable, S_x , is given by

$$S_x = 1/(N - 1) \sum (X_i - \bar{X})^2. \quad (2.4)$$

A simple example can be used to illustrate this point of why the sample variance must be adjusted by the degrees of freedom. One can always connect two points with a straight line. The mean value is the average of these two points. The variance is supposed to be the dispersion around the line connecting these two points, but there isn't any variance: the line connects the two points exactly, leaving no residual. Similarly, a plane can always be drawn through three points, and so on. The argument is the same as we move into n dimensions. The more variables that are contained in the equation, the more likely it is that the n -dimensional line will connect all the points, even if the relationship doesn't explain anything. Thus an unbiased estimate of the true variance must be calculated by adjusting for the degrees of freedom.

The square root of the sample period variance is known as the *standard deviation*, which is the more common measure used in statistical parlance. The comparison of the estimated mean to its standard deviation indicates whether that mean is statistically significantly different from some preassigned value, usually zero.

The mean and variance are the two sample statistics most often used to describe the characteristics of the underlying probability distributions. They are not the only ones. Statistics books generally refer to the methods of "moments," which show that the mean and variance are only the first and second moments of a long list of characteristics that describe various probability distributions. Sometimes it is useful to find out how much distributions deviate from the normal distribution by looking at the third and fourth moments, known as *skewness* and *kurtosis*. For example, a distribution might be "lopsided" with the peak value far away from the middle, which is skewness. The tails might be too "fat," which is kurtosis. Also, the distribution could have more than one peak. However, for practical purposes in most practical statistical work – including but not limited to economics – the mean and variance are the only tools that are used to describe the shape of the probability distribution. That is because the normal distribution, which is the most important distribution for statistical work, is completely defined by its mean and variance.

2.5 GOODNESS-OF-FIT STATISTICS

One of the major aims of this book is to explain how to build a forecasting model that will minimize forecast error. As will be seen in numerous examples, independent variables that appear to be highly correlated with the dependent variable in the sample period often show a much smaller correlation in the forecast period. Nonetheless, in a brief statistical review it is useful to indicate the tests used to determine which variables are statistically significant, and how well the equation fits, over the sample period. We want to determine if the parameter estimates – the coefficients – in the model are significantly different from zero, and also what proportion of the total variance of the dependent variable is explained by the regression equation. The statistical significance of each coefficient is determined by dividing the value of each coefficient by its standard error. If the residuals are normally distributed, the parameter estimates will generally be statistically significant from zero at the 95% probability level if this ratio is 2 or greater, and at the 99% level if this ratio is 2.7 or greater.

The proportion of the variance of the dependent variable explained by the equation is known as *R*-squared. It is sometimes thought that the higher the *R*-squared, the more accurate the forecasts will be; but as will be shown throughout this book, that is often not the case. Nonetheless, virtually every model builder looks at the values of *R*-squared in determining which equation to choose, and to a certain extent I will follow that general practice.

2.5.1 COVARIANCE AND CORRELATION COEFFICIENTS

We have defined the theoretical and sample mean and variance for each random variable *X*. However, from the viewpoint of statistics, econometrics, and forecasting, the interesting part is not so much the characteristics of a single random variable *X*, but its correlation with other random variables *Y* and *Z*. At this point we consider only the bivariate case, or the correlation between two random variables *X* and *Y*. To determine this correlation, we can calculate the *covariance*, which is defined as

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]. \quad (2.5)$$

Substituting the mean values of *X* and *Y* for their expected values, and switching from the true covariance to its sample period estimate, we have

$$\text{Cov}(X, Y) = \sum \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{N - 1}. \quad (2.6)$$

The *correlation coefficient* is defined as the covariance divided by the product of the standard deviation of *X* and *Y*. The point of this transformation is that the

size of the covariance depends on the scale factors used (millions, percent changes, square feet, etc.) whereas the correlation coefficient is always between -1 and $+1$, so one can see at a glance how strong the correlation is. The correlation coefficient is thus given as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \quad (2.7)$$

where σ_X and σ_Y are the standard deviations of X and Y respectively.

2.5.2 STANDARD ERRORS AND t -RATIOS

After determining the correlation coefficient, the model builder wants to know whether this correlation is significantly different from zero at some designated level, usually the 95% probability level. One could easily test whether the parameter estimate is significantly different from some other value, but most of the time researchers want to determine whether the coefficient is significantly different from zero.

Consider the simple bivariate linear equation

$$Y_t = \alpha + \beta X_t. \quad (2.8)$$

Estimating the regression equation yields an estimate of the intercept α and the slope coefficient β ; the least squares algorithm also supplies estimates of the variances of the estimated values of α and β . The significance level is determined by taking the ratio of the coefficient to its standard error, which is the square root of the variance. In everyday terms, this means the standard error serves as a measure of the dispersion of the probability distribution of that coefficient around its mean value. If the standard error is small relative to the coefficient, the probability is high that the actual value will be close to the estimate; if the standard error is large relative to the coefficient, the actual value could be just about anything, and the estimated value is not very useful. If the error term is normally distributed, then we can determine whether the coefficient is significantly different from some desired test value, generally zero.

Some actual numerical examples are provided later. For now, consider the case where the coefficient is 0.70 and the standard error is 0.30. Also assume that the error term is normally distributed. The ratio of the coefficient to the standard error is 2.33. What does that mean?

We have already noted that one rule of thumb – almost taken for granted in most of the empirical articles in economics – states that if this ratio is greater than 2, the variable is significantly different from zero; or, in short, significant. For the practicing econometrician, that is the rule used to show that your results are meaningful. Perhaps a better level of significance could be found, but this result is so ingrained in statistics that we will continue to use it.

As the sample gets smaller, the ratio of the coefficient to its standard error must be somewhat larger for any given level of significance. The ratio of the sample mean to the sample variance – as opposed to the population mean and variance – is known as a *t*-ratio. The *t*-distribution is similar to the normal distribution. In general it has fatter tails than the normal distribution, but approaches it as the sample size increases.

Tables for the *t*-ratio are given in any standard statistics or econometrics textbook. These tables show that as the sample size diminishes, the ratio of the coefficient to its standard error must be increasingly greater than 2 to be significant at the 5% level. Given that the 5% level of the normal distribution is 1.96 times the standard error, below are listed some values of the *t*-distribution to see how much difference the sample size makes. All these levels of significance are based on what are known as *two-tailed tests*; i.e., no a-priori guess is made about whether the sign ought to be positive or negative. If we knew for sure what the sign was supposed to be, the *t*-ratios would be about 20% lower for comparable levels of significance (e.g., the 5% level of significance would be 1.64 instead of around 2).

<i>Degrees of freedom</i>	<i>t-ratio for 5% significance</i>
5	2.57
10	2.23
15	2.13
20	2.09
40	2.02
60	2.00
∞	1.96

For practical purposes the difference narrows very quickly. As a general rule of thumb, this author suggests that you should not try to build a forecasting model using less than 20 independent observations. At that level, the difference between a *t*-ratio of 2.1 and 2.0 will probably be overwhelmed by other statistical difficulties in the data.

2.5.3 *F*-RATIOS AND ADJUSTED *R*-SQUARED

The *F*-ratio, which measures the overall significance of the estimated equation, can be defined as

$$F = \frac{X^*(n-k)}{Y^*(k-1)} \quad (2.9)$$

where *X* is the *explained* part of the variance of the dependent variable, and *Y* is the *unexplained* part. Also, *n* is the total number of observations and *k* is the number of estimated coefficients, so *n* – *k* is the number of degrees of freedom

in the equation, and $k - 1$ is the number of independent variables. The F -ratio can be used to test whether the explained part of the variance – compared with the unexplained part – is large enough to be significantly different from zero (or whatever other number is selected).

If the F -ratio measures the significance of the entire equation, and the t -ratio measures the significance of an individual coefficient, there ought to be some relationship between the two ratios for the bivariate case, which is

$$t^2 = F. \quad (2.10)$$

An intuitive explanation of this relationship is that the t -ratio measures the explained coefficient relative to its standard error, while the F -ratio measures the explained variance relative to the unexplained variance for the entire equation. In the bivariate case, t is the ratio of the explained part of the equation to the unexplained part, while F is the square of both those terms. Thus the F -ratio is usually considered only in multivariate equations; for the simple bivariate case, the F -ratio does not contain any additional information not already found in the t -ratio.

However, the F -ratio is not particularly easy to interpret without having the F -distribution tables in front of you. Does a value of 8.4 mean the equation is significant or not? (Answer: it depends on the number of degrees of freedom.) Recall that the covariance between two variables could be easily converted into a correlation coefficient that ranged between -1.00 and $+1.00$, which gave us an easy-to-interpret figure without further adjustment.

The F -ratio is amenable to a similar interpretation. The statistic most commonly used is known as R -bar squared, which is the proportion of the total variance of the dependent variable that is explained by the regression equation, adjusted for degrees of freedom. \bar{R}^2 is equally suitable for multiple regression equations, as will be seen in chapter 3. It is defined as

$$\bar{R}^2 = 1 - \frac{\text{unexplained variance} * (n - 1)}{\text{total variance} * (n - k)}. \quad (2.11)$$

This is similar to, but not exactly the same as

$$R^2 = \frac{\text{explained variance}}{\text{total variance}}. \quad (2.12)$$

To see the difference, suppose that the explained variance equals 95% of the total variance. Then R^2 would be 0.95. However, suppose there are 50 observations; then $n - 1 = 49$ and $n - k = 48$, so $\bar{R}^2 = 1.00 - 0.05 * (49/48)$, which is 0.949.

When n is large, R^2 is large, and k is small, there is very little difference between \bar{R}^2 and R^2 . However, as R^2 drops, the difference can be substantial, especially for small samples. In extreme cases, \bar{R}^2 can be negative. In this book, it is often listed as RSQ.

One word of caution: all these formulas are calculated by taking variables around their mean values. It is possible to calculate a regression equation without any constant term. In that case, the formulas do not apply and often give ridiculous values for R^2 that cannot be used; often the reported results are negative. Most programs will warn if you have inadvertently left out the constant term.

2.6 USING THE EViews STATISTICAL PACKAGE

The graphs shown in this text are produced by the EViews software program, which is used throughout this book. Just as there are hundreds if not thousands of sources of data, there are many different software programs written for the PC that can be used to estimate regressions and build models. However, for our purposes, the list can quickly be narrowed down to a few names.

The program should be primarily designed for economic model building, which means including an efficient simulation capability as well as estimating regression equations. It should be simple to generate transformations of the variables, including lags, percentage changes, and ratios, and it should also be easy to add dummy variables and estimate nonlinear equations. The program should also contain a full battery of standard tests to determine whether the various parameters are statistically significant. It should also permit easy data entry and exit and be compatible with existing large-scale economic databases. Other programs satisfying all these criteria include SAS, SPSS, PCGIVE, and RATS. Minitab and Excel are widely used for spreadsheet forecasting but are not so useful for building models. In this author's experience, the modeling capabilities of EViews are easier to use than those found in competing programs.

The examples and printouts in this text are based on EViews; other programs generally have similar formats and provide essentially the same information. Figure 2.1 shows a typical printout, and the following text identifies some of the standard terms that appear along with each regression equation to show the reader what is expected. For most of the equations in this book, an abbreviated form is used to convey the most important statistical information.

- @PCH(WAGERATE) is the dependent variable. The symbol WAGERATE stands for an index of average annual wage rates. @PCH means percentage changes are being used. In EViews, percentage changes are *not* multiplied by 100, so a change from 1.00 to 1.05 would appear as 0.05 rather than 5.0.
 - The sample period is given along with the number of observations, in case any years were skipped because of missing data. In this case, there are 50 years from 1949 through 1998 inclusive, so no data are missing. From time to time it might be advisable to omit one or more observations if it appeared to be far out of line with the rest of the information. Alternatively, data might be missing for one or more observations.
-

Dependent Variable: @PCH(WAGERATE)
Method: Least Squares

Sample(adjusted): 1949 1998

Included observations: 50 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.009	0.003	2.94	0.005
@PCH(CPI)	0.613	0.049	12.56	0.000
@PCH(MINWAGE)	0.042	0.008	5.01	0.000
@PCH(POIL)	-0.013	0.005	-2.66	0.011
@PCH(M2,2)	0.078	0.020	4.01	0.000
1/UN(-1)*DBR	0.079	0.010	7.67	0.000
R-squared	0.90	Mean dependent var		0.056
Adjusted R-squared	0.89	S.D. dependent var		0.023
S.E. of regression	0.0076	Akaike info criterion		-6.80
Sum squared resid	0.0027	Schwarz criterion		-6.57
Log likelihood	176	F-statistic		80.3
Durbin-Watson stat	2.07	Prob(F-statistic)		0.000

Figure 2.1 A typical output of EViews. The elements are explained in the text.

- @PCH(CPI) is the percentage change in the consumer price index.
- @PCH(MINWAGE) is the percentage change in the minimum wage.
- @PCH(POIL) is the percentage change in the price of crude oil. This enters with a negative sign to show that when there are major swings in oil prices, wage rates do not adjust as much as when changes occur in the CPI due to other factors.
- @PCH(M2,2) is the percentage change in the M2 measure of the money supply over the past two years.
- UN is the unemployment rate. It used to be thought that when the unemployment rate declined, wage rates increased. However, once Paul Volcker reestablished the credibility of monetary policy in 1982, that term was no longer needed, so it is zeroed out starting in 1982. The reader can verify that adding a term $1/UN(-1)*(1 - DBR)$ – where DBR is 1 before 1981 and 0 afterwards – has a t -ratio that is very close to zero.
- The coefficient for each term (including the constant term) is followed by its standard error. The t -statistic is the ratio of the coefficient to its standard error. The “prob” column shows the probability that the coefficient is not significantly different from zero. For the percentage change of oil term, the probability is 0.011 that the term is zero. The CPI is clearly quite significant; the probability that it is zero is less than 0.00005.

- R-squared is the percentage of the variance of the dependent variable explained by the equation. Adjusted R-squared (often called RSQ in this text) is R^2 adjusted for degrees of freedom; in this case it is 0.89. The standard error of the regression equation is 0.0076, or 0.76%. That means that approximately two times out of three, the sample period error for predicting the wage rate was less than 0.76%, compared with an average change of 5.6% (noted below). The sum of squares residual is the standard error squared multiplied by the degrees of freedom; it does not add very much information.
- The log likelihood ratio is used to test for maximum likelihood estimates, which are not considered in this book, and can be ignored here.
- The Durbin–Watson statistic is discussed in chapter 3. It is a test for the autocorrelation of the residuals. If no autocorrelation is present, the DW statistic is 2. If this statistic is less than about 1.4, the residuals are serially correlated. When that happens, the t -ratios and R^2 are overstated, so the equation will usually not predict as well as indicated by the sample period statistics. The DW of 2.07 indicates there is no autocorrelation in this equation.
- The mean dependent variable is 0.056, which means the average annual change in wage rates over the sample period was 5.6%. The SD dependent variable line show the standard error of the variable around its mean, which is 2.3%.
- The Akaike and Schwarz criteria are designed to show whether an equation would be better by containing fewer terms; those are used for time-series models and are discussed in Chapter 7.
- The F -statistic measures where the overall equation is significant; the probability that the entire relationship is not statistically different from zero is 0.000000. Since several terms are significant, the overall equation must be significant in any case; for this reason, the F -ratio is not used very often. While it is possible to estimate an equation where none of the t -ratios is greater than 2 but the overall F -statistic was significant, that would mean cluttering the equation with individual terms that are not significant, which would ordinarily generate a very poor forecasting equation.

A brief note on the number of significant digits. The actual program for EViews generally shows six or seven numbers. I have reduced this clutter to show two or three significant figures, which makes more sense economically. There is no real difference between, say, \bar{R}^2 of 0.8732 and 0.8741, or between t -ratios of 5.88 and 5.84.

A typical graph, showing the actual values, those estimated by the equation, and the residuals, is in figure 2.2. The top half of this figure shows the actual values of changes in wage rates compared with the estimated values calculated by the regression equation in figure 2.1; these are also called simulated or fitted values. The bottom half shows the residuals, defined as the actual minus the fitted values. The largest error occurs in 1989, when wages rose far less than would be predicted by the equation; almost as large a discrepancy occurred in 1992 in the other direction.

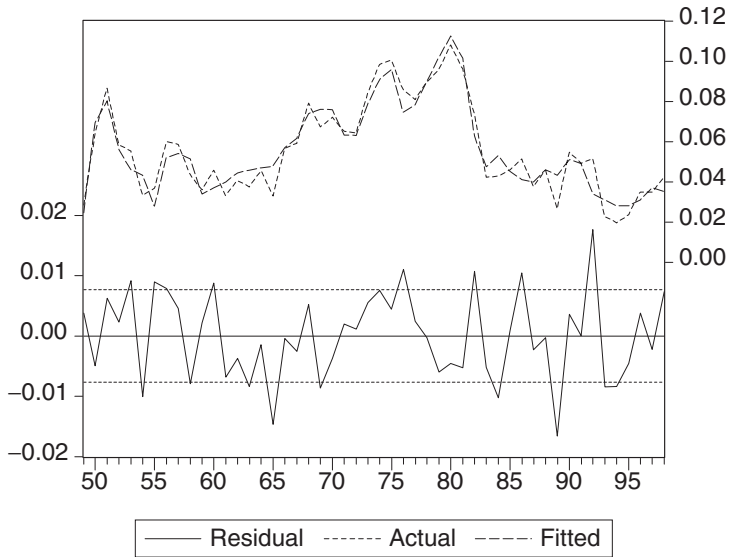


Figure 2.2 A typical output from EViews. See the text.

2.7 UTILIZING GRAPHS AND CHARTS

The construction of econometric models is often based on economic theory. However, in virtually all cases, the researcher looks at the underlying data in order to form some opinion of how the variables are correlated, and whether the correlation is improved when the independent variables are lagged.

There are three principal methods of displaying time-series data. Line graphs usually show two or more series graphed against time. Scatter diagrams have all the sample period points for one variable on the y -axis and the other variable on the x -axis. Bar graphs are often utilized to describe the characteristics of a single series; the most common use in this text is histograms, where either the original series or the residuals from a regression equation can be checked for normality and other statistical properties. Bar graphs can be used for multiple variables, either on a side-to-side basis or stacked. Sometimes pie charts are used as graphical aids, but these are usually for a snapshot of events at some given time and are not ordinarily utilized with time-series data.

The well-known comment about lies, damn lies, and statistics, variously attributed to Benjamin Disraeli and Mark Twain among others, summarizes how many people view graphical analysis. The same data can tell completely different stories depending on how they are presented. To see this, consider the simple relationship between consumption and disposable income, both in constant dollars. Figure 2.3 shows a line diagram of the difference between actual and simulated consumption. It looks like almost a perfect fit. Figure 2.4 shows

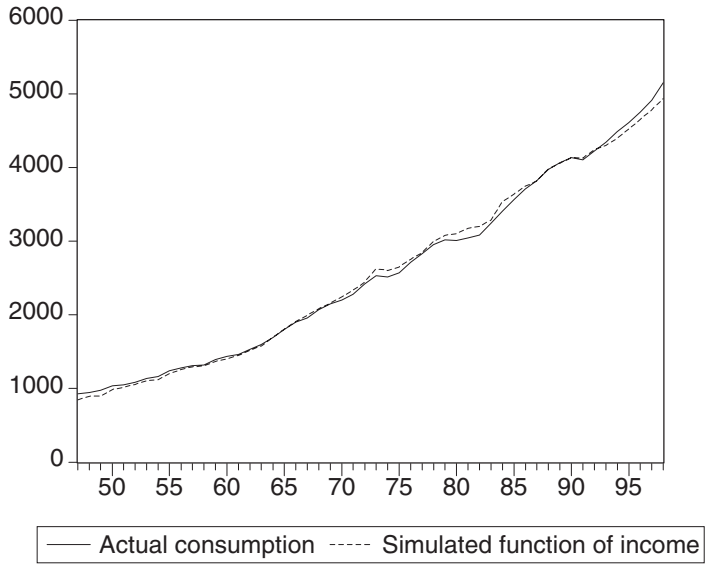


Figure 2.3 The level of real consumer spending appears to follow real disposable income very closely.

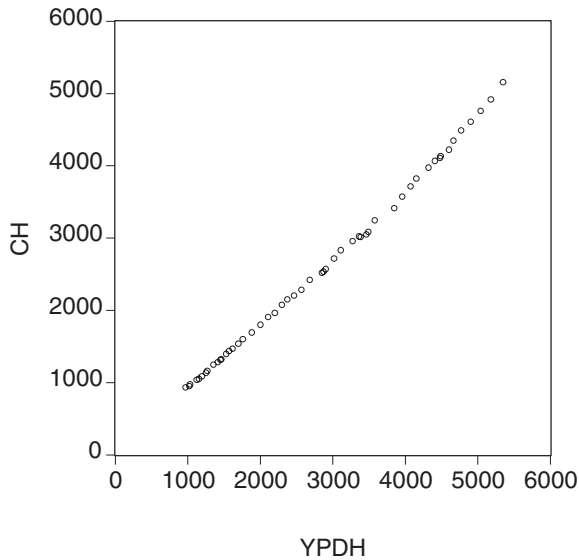


Figure 2.4 The scatter diagram between consumption and income shows almost a perfect fit.

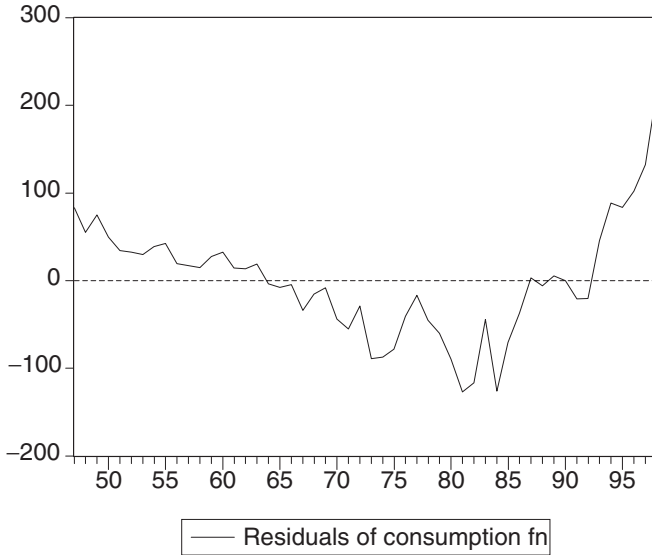


Figure 2.5 The residuals when consumption is regressed on real disposal income are quite large in many years.

the same data in a scatter diagram, which reinforces that conclusion. Yet figure 2.5 shows the residuals of that equation; when put on a different scale, it is more easily seen that the errors in predicting consumption with this simple equation may be as much as \$200 billion per year.

One could claim that, without reference to further benchmarks, we don't know whether \$200 billion is a "large" error or not. Some further comparison is warranted. In 1999, real consumer spending in the US was about \$6,000 billion, and over the past 10 years had grown at an average annual rate of 3.5% per year. Hence a naive model that said the growth rate in 2000 would continue at 3.5% would predict an increase of about \$210 billion. In fact the actual increase, based on preliminary data, was \$316 billion, for an error of \$106 billion. Seen in that light, a \$200 billion error is abnormally large, since it is almost double the error generated by a naive model.

Finally, figure 2.6 shows the actual and forecast values for the percentage changes in each of these variables; which makes it obvious that while income is an important determinant of consumption, it is hardly the only one. The lines in the top part of this graph show the actual percentage change in consumption compared with the percentage changes that are estimated by the regression equation, which in this case simply states that percentage changes in consumption are a function of percentage changes in income plus a constant term. The line in the bottom part of this graph, which is on a different scale, plots the residuals, or the differences between the actual and estimated values of the dependent variable.

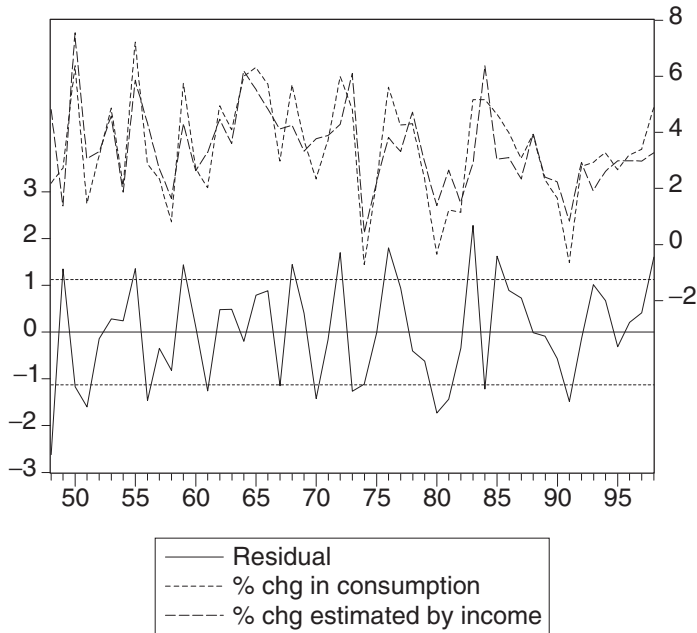


Figure 2.6 The percentage change in income is an important determinant of the percentage change in consumption, but the residuals are still quite large.

In another example, consider the correlation between the Federal funds rate and the rate of inflation, as measured by the consumer price index (CPI), on an annual basis. In general, we see that when the inflation rate changes, the Federal funds rate is likely to change by a similar proportion.

Figure 2.7 shows a scatter diagram with annual data for the funds rate and the inflation rate for the period 1955 through 1998 (no data are available for the funds rate before 1955). It is clear the series are positively correlated, although not perfectly. The solid line represents the regression line as calculated by least squares. Note that the slope of the regression line is slightly less than unity, which means when the inflation rate is zero, the funds rate is slightly positive.

Figure 2.8 shows the same two variables using a line graph. From 1955 through 1980, the funds rate exceeded the inflation rate by only a small amount. From 1981 through 1989, the gap between the funds rate and the inflation rate was much greater, indicating a shift in Federal Reserve policy. The line graph shows this clearly, whereas the scatter diagram does not.

According to the assumptions of the classical linear model, the residuals are supposed to be normally distributed. One simple test is to examine the histogram of the residuals to see whether that is indeed the case. We look at the residuals from the equation shown above, where the Federal funds rate is a function of the inflation rate.

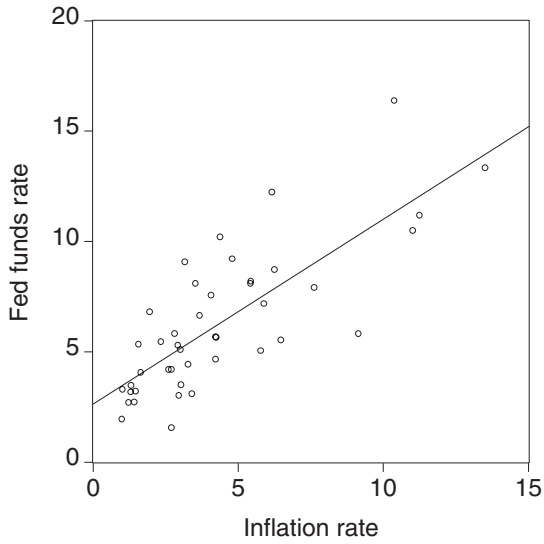


Figure 2.7 When inflation rises, the Fed funds rate also increases, but not quite as rapidly.

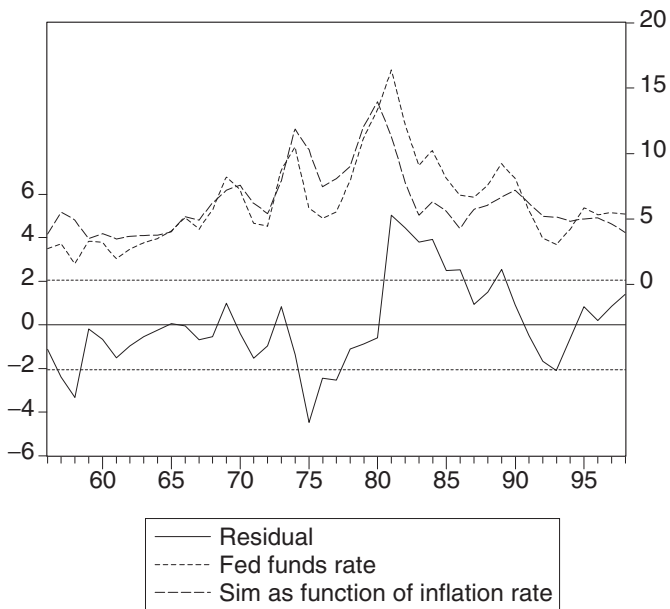


Figure 2.8 After 1980, the Fed funds rate was usually much higher than the inflation rate.

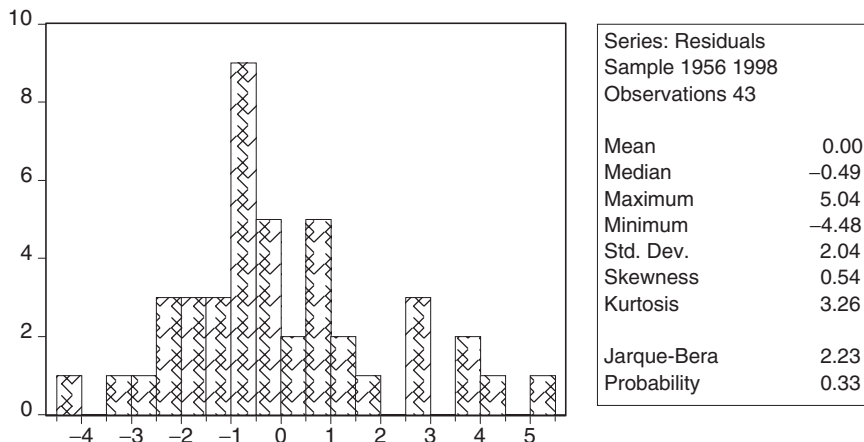


Figure 2.9 When annual data are used, the residuals from the equation where the Fed funds rate is a function of the inflation rate are normally distributed.

The residuals from the equation using annual data are normally distributed, as shown in figure 2.9. The graph, which is taken from EViews, is accompanied by several statistics. By definition the mean is zero. The median is slightly negative, indicating that there are more negative than positive residuals. The maximum and minimum values of the residuals are given next. The standard deviation is 2.03.

The next line measures skewness, which is the measure of how much the distribution is lopsided. If this were a perfectly normal distribution, skewness would be zero. Kurtosis measures the “fatness” of the tails; for the normal distribution, kurtosis is 3. A casual glance indicates the calculated measures of skewness and kurtosis are not very far away from the values of a normal distribution, but we need a formal test. The standard measure is known as the Jarque–Bera (JB) statistic, which is defined as

$$JB = \frac{(N - k) * \left[S^2 + \frac{(K - 3)^2}{4} \right]}{6} \quad (2.13)$$

where N = number of observations, k = number of variables in the equation, S = skewness, and K = kurtosis. The probability 0.33 means that one would observe a JB statistic this high 33 percent of the time under the hypothesis that the residuals are normally distributed. Since that is well above the usual 5% level of significance, in this particular case the residuals are normally distributed.

However, if we run a regression with the same variables using quarterly instead of annual data, a different result emerges for the residuals. As shown in

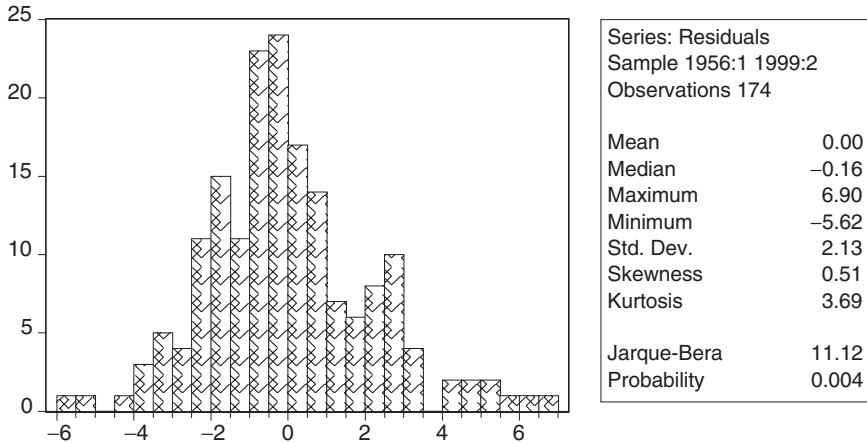


Figure 2.10 When quarterly data are used, the residuals from the equation where the Fed rate is a function of the inflation rate are not normally distributed.

figure 2.10, kurtosis is much higher, as shown by the proliferation of outlying values with both positive and negative signs. As we will see later, the main reason is that, on a quarterly basis, the Federal funds rate depends on the lagged as well as current values of inflation. The point illustrated here is that using annual and quarterly data can give far different statistical results even if the coefficients are quite similar.

2.8 CHECKLIST BEFORE ANALYZING DATA

When teaching courses in forecasting, I have found that one of the most frustrating tasks is to convince students to check the data before they start using them. Even if the data are obtained from reputable sources, mistakes happen. Sometimes the series are corrupted, and sometimes the starting and ending dates are not exactly as listed. Even if the data are error-free, one or two outliers may distort the entire model-building process; unless you check ahead of time, that won't become apparent until your regression estimates provide unrealistic sample period estimates or inadequate forecasts. Sometimes series that are supposed to be in comparable units are not; one series is in millions, while the other is in thousands.

Except for financial markets, most government data are seasonally adjusted, but most company data are not. Thus if you are going to mix the two types of data, some adjustment procedure is required. This topic will be discussed more in Part III, but at this juncture we look briefly at some of the major seasonal adjustment methods, including their plusses and minuses.

2.8.1 ADJUSTING FOR SEASONAL FACTORS

Most economic time-series data have seasonal patterns. For the most part, government data have already been seasonally adjusted, but this is not usually the case for individual company data. Attempts to use these data for modeling efforts without first applying seasonal factors will usually lead to suboptimal results.

Typical examples of seasonal patterns in economic data are the following: sales rise every Christmas, more people visit the beach in the summer, sales of snow shovels rise every winter, broiler (chicken) prices peak in the week of July 4, the unemployment rate for construction workers rises in the winter, and so on. To the extent that these patterns are regular, it is best to remove the common seasonal factors; otherwise one could end up with a correlation based on seasonal factors rather than underlying economic trends. The classic story here is about the economist who correlated seasonally unadjusted consumer spending with unadjusted money supply figures; since both of them rise sharply in the fourth quarter, a spuriously high correlation was obtained. Some wag suggested this economist had “discovered that the money supply causes Christmas.”

Suppose one calculated a regression for unseasonally adjusted department store sales on dummy variables for each month of the year (e.g., the variable for December would be 1 for that month and 0 elsewhere, and so on). That regression would produce a very high correlation, but the equation would have explained nothing except that department store sales data rise before Christmas and Easter and fall during February and July. The fit would be high, but an equation of that sort would contain no relevant information. What retailers usually want to know is whether sales this year – or this Christmas season – will be better or worse than usual, adjusted for the overall growth trend.

After removing the trend and seasonal factors, the data series that remain is more likely to resemble a random variable and hence more closely satisfy the basic statistical criteria and tests. As a result, the statistical results that are obtained are more likely to provide a realistic appraisal of how accurate the forecasts will be. Of course that does not guarantee that the results will be useful, but it does improve the odds.

2.8.2 CHECKING FOR OUTLYING VALUES

Once the data have been successfully entered into EViews or a similar program, it is quite simple to create a histogram for each variable and make sure that outlying observations will not dominate any regression equations that might be estimated. Take the time; it's well worth it.

Technically, only the residuals need to be normally distributed to satisfy the usual statistical criteria. However, if there are outliers you should either exclude them or treat them with dummy variables; otherwise they will dominate the regression results. Later I show what happens when outliers are ignored.

Suppose an observation is five standard deviations from the mean. If the variable really is normally distributed, the odds of that occurring are only about one in a million. Yet as a practical matter, since the sum of squares is being minimized, such a residual would have a weight 25 times as great as an observation that is one standard deviation from the mean. In a modest sample size of 20 to 50 observations, that one outlier would dominate the regression equation and in effect the regression would just be fitting that point.

Figure 2.11 shows the histogram of quarterly percentage changes in auto sales. Clearly the changes are not normally distributed. There is substantial kurtosis (fat tails), and the probability that this series is normally distributed is less than 10^{-6} . Given that fact, the next question is whether there is any compelling economic reason for those outliers. To answer that question, we turn to a time-series plot of the data, which is shown in figure 2.12.

It is clear that the major pairs of outlying observations occurred in 1959.4/60.1, 1964.4/65.1, and 1970.4/71.1. The first pair was caused by a major steel strike; the others were major auto strikes. Thus strike periods should be handled differently. In this case a dummy variable for auto strikes is the most appropriate treatment; in other cases, outliers should be omitted entirely.

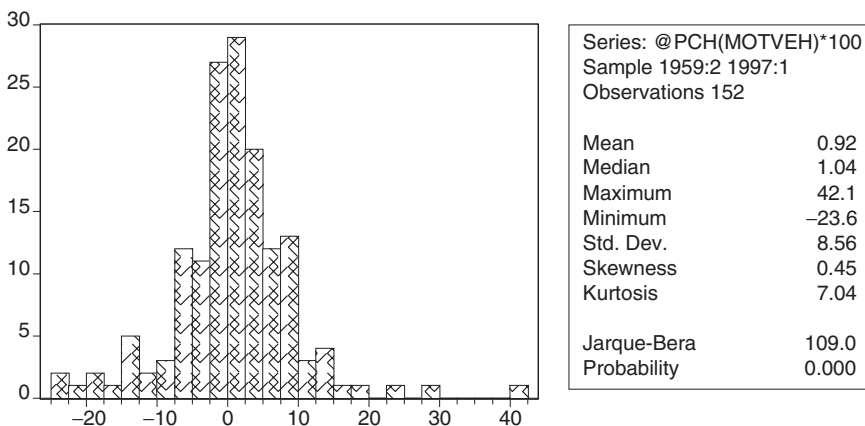


Figure 2.11 Histogram of percentage changes in quarterly motor vehicle sales.

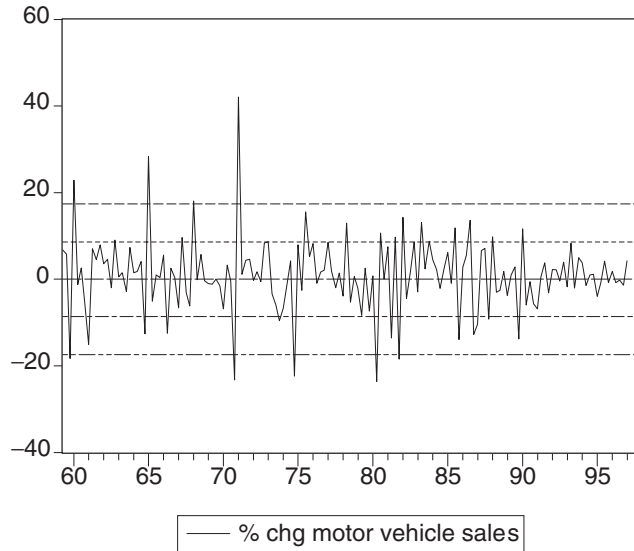


Figure 2.12 Percentage change in quarterly motor vehicle sales. The horizontal lines are 2, 1, 0, -1, and -2 standard deviations from the mean.

2.9 USING LOGARITHMS AND ELASTICITIES

One of the key themes in this book is that model builders should eliminate spurious trends by a variety of methods, including percentage first-difference equations, in equations where two or more variables have strong time trends. Also, there are often many cases where logarithms should be used, particularly if the underlying theory suggests a constant elasticity for the parameter being estimated. The use of logarithms often reduces the spurious effect of common upward trends, while using logarithms instead of percentage changes reduces the chances of one or two extreme values distorting the entire equation. Since the relationship between coefficients and elasticities is sometimes confusing, it is briefly reviewed.

The next two figures show the historical pattern of the S&P 500 stock price index in levels and in logarithms. Figure 2.13 seems to indicate that the market is rising at ever-more rapid rates, but in fact that is not the case. Figure 2.14 shows that from 1947 through 2000, this stock price index has advanced about 7% per year; it rose less rapidly during the period of high interest rates in the late 1970s and early 1980s, and more rapidly in the late 1990s and 2000, when it appeared to some investors that inflation and interest rates had moved to “permanently” lower levels. Except for these diversions, the long-run growth rate of stock prices is seen to be quite steady.

An *elasticity* measures the percentage change of a given variable relative to some other variable. Suppose that a 1% increase in the price of food results in

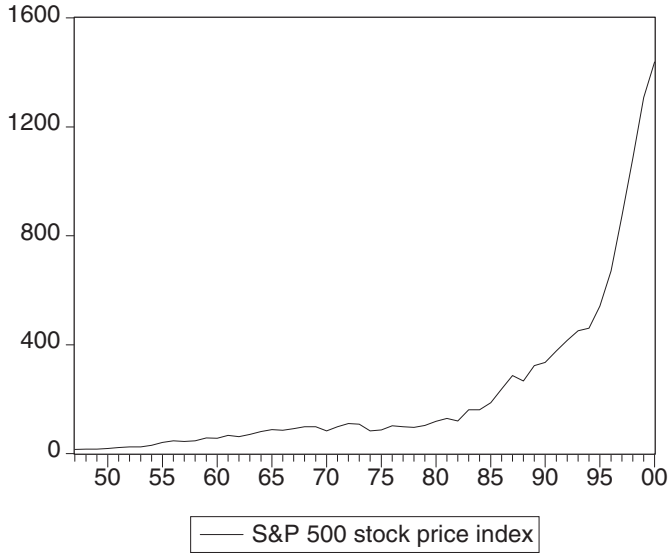


Figure 2.13 Using levels, the S&P 500 stock price index seems to be increasing at an ever-faster rate.

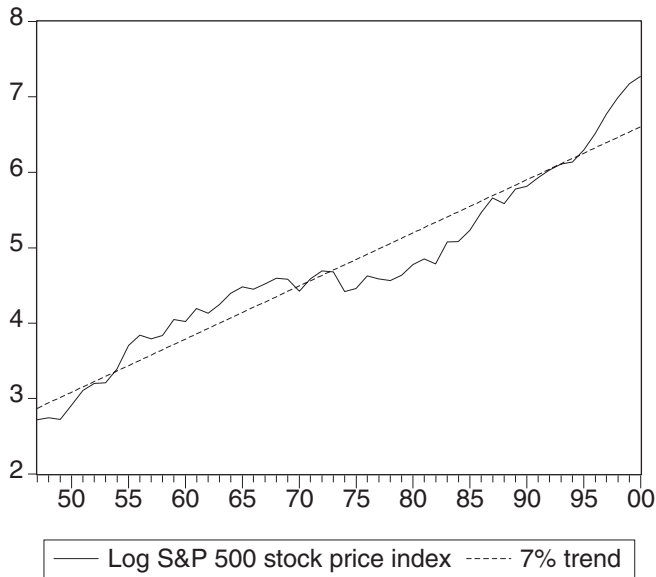


Figure 2.14 The logarithmic version of figure 2.13.

a 0.4% decline in purchases of food, *ceteris paribus*. In that case, the price elasticity of food is -0.4 .

Both logarithms and elasticities measure percentage changes. Furthermore, in the regression equation $\log y = a + b \log x$, the coefficient b is the elasticity of y with respect to x . Hence estimating this equation in logarithms assumes the elasticity remains constant over the sample period. Using logs provides a convenient measure of the elasticity being estimated.

Because of how logarithms are defined, $\log x - \log x_{-1}$ is approximately equal to $(x - x_{-1})/x_{-1}$ and $\log y$ is approximately equal to $(y - y_{-1})/y_{-1}$. That means, as a first approximation:

$$\log y - \log y_{-1} = b(\log x - \log x_{-1}) \quad (2.14)$$

can be written as

$$(y - y_{-1})/y_{-1} = b(x - x_{-1})/x_{-1}. \quad (2.15)$$

If we take the definition of elasticities at their mean value, then

$$\eta_{yx} = \frac{(y - y_{-1})/y_{-1}}{(x - x_{-1})/x_{-1}} \quad (2.16)$$

so that b and η_{yx} are the same. In a similar vein, equations that compare percentage change of levels and first differences of logarithms will give almost identical results.

Problems and Questions

1. Use the data for monthly stock prices as measured by the S&P 500 (all necessary data can be collected from the website).
 - (a) Calculate the mean and variance for this series.
 - (b) Now take the first difference of this series and recalculate the mean and variance, and the percentage first difference and recalculate the mean and variance.
 - (c) Calculate a simple regression where stock prices are a function of a time trend. Calculate the variance of the residuals of this equation.
 - (d) Which of these four methods has the smallest variance? What meaning, if any, does that have for forecasting the stock market one month from now? Five years from now?

continued

2. Run a least squares regression where constant-dollar (real) consumption is a function of constant-dollar disposable income for the period 1947–92.

- (a) Show your results. What is the standard error of estimate for that period?
- (b) Now use this equation to “forecast” the values of real consumption for 1993–8 using the same equation. What is the standard error of the forecast for those six years? How does it compare to the standard error during the sample period?
- (c) What factors do you think made the standard error so much larger in the forecast period?
- (d) Now run a regression where the percentage change in real consumption is a function of the percentage change in real disposable income. Compare the R^2 and standard errors from these two equations.
- (e) Use the percentage change equation to predict consumption for the 1993–8 period. Compare the standard error during this period with the standard error from the levels equation. Which equation is better for forecasting?

3. Plot the monthly percentage changes in the PPI-finished goods index.

- (a) Test this series for normality, using a histogram. What do the results show?
- (b) Use the same series, but use 12-month percentage changes. Using the same histogram analysis, are the residuals now normally distributed? Why or why not?
- (c) Calculate a regression of the 12-month percentage changes in the PPI on 12-month percentage changes in the price of oil. Examine the residuals. What does this suggest about the lag structure? Try using lagged values of the percentage changes in oil prices to determine the optimally fitting lag structure, and show your results.
- (d) Explain why you would or would not use this equation for forecasting changes in the PPI. Take into account the DW statistic and the normality tests in your answer.

4. Plot a scatter diagram of non-farm payroll employment and real GDP using quarterly data.

- (a) Based on this diagram, what would you say about the relationship of these two variables?

continued

- (b) Now plot a scatter diagram of the percentage changes in the same two variables. Do you still think they are closely correlated?
 - (c) Calculate a regression of employment on real GDP with no lag. Now expand the equation to include lags of one and two quarters for GDP. How do the coefficients change? Do you think this is a sensible interpretation?
 - (d) Rerun this equation using logarithms, with log of real GDP lagged by zero, one, and two quarters. Compare and contrast the coefficients.
 - (e) Now calculate a regression where the percentage change in employment is a function of the percentage change in GDP with lags of zero, one, and two quarters. Of these three equations, which one would probably be the most useful for forecasting employment? What criteria did you use to make this decision?
- 5.** Monetarists often claim that the rate of inflation is closely tied to changes in the money supply lagged two years.
- (a) Create a scatter diagram between the inflation rate and changes in M2 lagged two years for the sample period 1947–2001. Do you think that correlation is very strong?
 - (b) Now subdivide the sample period into two periods, 1947–81 and 1982–2001. Based on the scatter diagrams, during which period is the correlation with the money supply stronger? What happened in 1982 to change this relationship?
 - (c) Now plot a scatter diagram using the rate of inflation and the change in unit labor costs. The correlation is much stronger. Does that necessarily mean the latter equation would provide better forecasts of inflation? (Hint: how difficult is it to predict unlagged unit labor costs?)
- 6.** Regress the log of constant-dollar tobacco shipments on the relative price of tobacco.
- (a) What is the price elasticity, according to this equation?
 - (b) Now add a time trend to that equation. What happens to the elasticity?
 - (c) Now regress the percentage change in real tobacco shipments on the percentage change in the relative price of tobacco. What is the coefficient, and how does that compare with the elasticity?
 - (d) Several sophisticated studies have shown that the price elasticity of tobacco is approximately -0.45 . How do these results compare with the simple calculations in (a)–(c)?

continued

(e) By shortening the sample period, observe what happens to the elasticity. Based on that finding, what do you think will happen to the demand for cigarettes as the relative price continues to rise?

7. Calculate a regression of current-dollar purchases of computers on a time trend.

(a) Now regress the log of computer purchases on a time trend. According to this equation, what is the average annual growth rate of computer sales?

(b) Based on this regression, what do you think will happen to the average annual growth rate in computer sales over the next decade?

8. Using the data from the latest *Wall Street Journal*, plot the yield curve for Treasury securities.

(a) Using historical quarterly data, plot the three-month Treasury bill rate, two-year note rate, 10-year note rate and 30-year bond rate (which does not begin until 1977.2). Most of the time, the longer the maturity, the higher the rate. However, there are several occasions when that does not occur. Identify those time periods.

(b) Now plot the Federal funds rate and the Aaa corporate bond yield (data begin in 1955.1). Note the periods when the funds rate is higher than the bond yield. Are they the same as the times when the Treasury yield curve was inverted?

(c) Plot the percentage change in real GDP with the *yield spread* between the 10-year Treasury note rate and the three-month bill rate lagged one year. What happens to real GDP every time the yield spread turns negative. Do you think that would be a valuable tool for predicting whether there will be a recession next year?

(d) Now repeat the plot for (c) using annual data. Has your conclusion reached in (c) changed?

(e) As a practical business forecaster, why do you think economists have been unable to predict any recessions in the past 40 years including the 2001 recession, even though each one has been preceded by an inverted yield curve? (Hint: what happens to forecasters who are wrong when everyone else is right, compared to those who make the “consensus” error?)

