Practical Business Forecasting
Michael K. Evans

CHAPTER 1

CHOOSING THE RIGHT TYPE OF FORECASTING MODEL

INTRODUCTION

Practical business forecasting is both a science and an art. It is a science in the sense that correct use of sophisticated statistical tools will invariably improve forecasting accuracy. It is an art in the sense that empirical data seldom if ever provide an unequivocal answer, so the user must choose between alternative relationships to select those equations that will provide the most accurate forecasts.

There are no perfect forecasts; they always contain some error. While perhaps that is obvious, it is nonetheless important to emphasize this fact at the outset. The point of this book is to show how to minimize forecast error, not to pretend that it can be eliminated completely. To accomplish this goal, a variety of forecasting methods may be used. In many cases, these methods will be complementary, not competitive.

Forecasts can be used for many purposes. Sometimes, predicting the direction of change is sufficient. For example, a model that could accurately predict the direction of the stock market the following day – even without providing any information about how much it would rise or fall – would be extremely valuable and profitable. No such model has ever been successfully constructed, although many have tried, and the goal will presumably remain elusive. At the other extreme, a model that predicted the direction of change in the consumer price index (CPI) the following month without forecasting the magnitude would be virtually useless, since over the past 40 years the monthly changes in the CPI have been negative only about 1 percent of the time.

There are many ways of forecasting, not all of which are based on rigorous statistical techniques. In some cases, informed judgment can provide the best forecasts, such as when "insiders" have company information that is not available to anyone else. Surveys may provide useful information about forecasts for the overall economy, specific sectors, or individual industries and firms. To the extent that these methods improve forecasting accuracy, they should be utilized.

Nonetheless, there is no rigorous way of testing how much informed judgments or survey techniques have boosted forecast accuracy, so they are mentioned only peripherally in this text. Instead, this text concentrates on illustrating how statistical and econometric methods can be used to construct forecasting models and minimize forecast errors. Initially, most economic forecasts were generated with structural equations; more recently, time-series analysis has been utilized more effectively. The benefits and shortcomings of both methods for generating optimal forecasts are identified.

This book is not a theoretical text; the emphasis is placed on practical business forecasting. As a result, theorems and proofs, which can be found in many other texts, will be kept to a minimum, with most of the material related to actual forecasting examples. In particular, this text will illustrate how statistical theory often needs to be adjusted to take into account those problems that recur in actual empirical estimation. Methods of adjusting the models to increase predictive accuracy are not to be denigrated or dismissed; they are an integral part of practical business forecasting.

1.1 Statistics, Econometrics, and Forecasting

Statistics is the application of probability theory and other mathematical methods to improve decision making where uncertainty is involved. Statistical theory and results are used widely in economics, but also apply to a large and diverse number of other disciplines, including sociology, agriculture, astronomy, biology, and physics.

The use of statistics is designed to provide answers where uncertainty exists, whether the uncertainty is due to randomness, or ignorance about the true underlying relationship that is to be tested. To illustrate the first case, we know the underlying probability distribution and hence the proportion of straights that will be dealt in a poker hand over the long run, but not what the next hand will show. To illustrate the second case, we probably do not know the true underlying relationship between capital spending and the rate of interest, or between the rate of inflation and the rate of unemployment, or changes in the value of the dollar and net exports. Cases where the underlying probability distribution is known are rare in economics.

Econometrics is the application of statistical and mathematical methods to the analysis of economic data to verify or refute economic theories. When structural equations are used, a specific theory is being tested for verification or rejection. By comparison, statistical methods are increasingly used with economic data to obtain parameter estimates without specifying any particular theory. Those models are usually known as time-series analysis; one standard technique is integrated autoregressive moving-average (ARIMA) models. Those models consist of correlating a given economic variable with its own lagged values, adjusted for trend and seasonal factors; no attempt is made to postulate an underlying theory.

Economic forecasting often relies on statistical or econometric methods, but even that need not be the case. Some types of forecasts do not involve mathematical techniques at all; for example, surveys or polls may produce valuable forecasts without utilizing any econometric methods. However, these types of forecasts are not featured in this book. Most of the examples will be confined to those types of forecasts that use statistical methods.

1.2 The Concept of Forecasting Accuracy: Compared to What?

No forecast is ever perfect; opinions about what will happen in the future invariably contain errors. Anyone who has ever attempted to predict anything knows that. On the other hand, forecasting can be quite useful if it provides better answers than alternative methods of guessing about the future. The relevant test for any forecast, then, is never whether the results contain errors, but how accurate they are compared to the alternatives. Like that old Henny Youngman one-liner "How's your wife?" the appropriate answer is always "Compared to what?"

Throughout this book, the difference between the *science* of statistics and econometrics and the *art* of forecasting is emphasized. Most of the sophisticated theorems and proofs in those fields are based on highly unlikely assumptions about the distribution of the error terms, and furthermore assume that the data generating process remains the same in the sample and the forecast periods. Adjusting models to generate better forecasts when these assumptions are not satisfied has often been disdainfully called ad hoc adjustment, unworthy of the name of econometrics. Yet it plays a vital role in improving forecast accuracy.

From 1940 through 1970, primary emphasis was placed on theoretical refinements of statistical and econometric procedures, with scant attention paid to systematic methods of adjusting forecasts to improve their accuracy. When macroeconomic models proved unable to predict any of the major changes in the economy in the 1970s, the emphasis gradually shifted to developing methods that produced useful forecasts even if they did not follow the theoretical procedures developed in earlier decades.

In *Forecasting Economic Time Series*, a reference book recommended for those with more advanced mathematical skills, Clements and Hendry[1] have classified the basic issues in forecasting, as opposed to econometrics. They state that "The

[1] Clements, Michael P., and David F. Hendry, *Forecasting Economic Time Series* (Cambridge University Press, Cambridge, UK), 1998.

features of the real-world forecasting venture that give rise to the sources of forecast error . . . induce a fairly radical departure from the literature on 'optimal' forecasting . . . and at the same time help to explain why some apparently ad hoc procedures work in practice" (p. 3).

The approach used here is much less mathematically rigorous than Clements and Hendry's. Also, the discussion of forecasting accuracy begins with structural models and then moves to time-series analysis, contrary to the procedure that they (and others) use. Yet the methodology in which real-world practical forecasting is approached is very much in the spirit of their approach. While the method of least squares is used for the vast majority of the examples, the reader should always keep in mind that the assumptions of the classical linear model seldom hold in practical business forecasting.

In many cases, the underlying data generating function has shifted, the variables are not normally distributed, the residuals are not independent, and the independent variables are not known at the time of the forecast. Even more important, repeated rerunning of regression equations, and the substitution of different empirical data series that measure the same theoretical concept, often help to improve forecasting accuracy but are outside the constructs of the classical linear model. For this reason, the statistical estimates generated under those assumptions must be interpreted carefully, and often with a degree of skepticism.

It is too crude to say that what makes the best forecasts is "whatever works," but that is closer to the spirit of the present approach than the method of choosing rigorous statistical techniques that minimize the root mean square error or other similar measures in the sample period but generate suboptimal forecasts. Sometimes structural econometric models provide better forecasts, and sometimes the results from ARIMA models with no economic structure are better. In certain cases, the *combination* of these methods will generate better forecasts than either method separately. Far from being relegated to the criticism of ad hoc adjustments, changing the model during the forecast period will invariably produce better results than a "pure" unadjusted model, provided it is done properly.

As Newbold and Granger have written,[2] "the evaluation criteria employed should be as demanding as possible since the object ought to be self-criticism rather than self-congratulation" (p. 266). The principal aim should be to build a forecasting model that will generate the smallest forecasting error, not necessarily maximize the goodness-of-fit statistics over the sample period.

The reader should always keep in mind that any forecasting model, no matter how sophisticated the underlying statistical techniques, must perform better than forecasts generated by random variables or naive methods. That means

[2] Granger, C. W. J., and Paul Newbold, *Forecasting Economic Time Series* (Academic Press, San Diego, CA), 1986.

always checking whether the model provides better results than other methods that are available – including naive models, surveys, and qualitative judgments.

A naive model generally assumes that the level or rate of change of the variable to be predicted this period will be the same as last period, or the change this period will be the same as the average change over an extended time period. For a time series without any significant trend, such as the Treasury bill rate, a naive model might state that the bill rate this month will be the same as it was last month. For a time series with a significant trend, the naive model would usually be couched in terms of percentage changes. For example, a naive model might state that the percentage change in the S&P 500 stock prices index next month will equal the percentage change last month, or it might equal the average percentage change over the past 480 months. A more sophisticated type of non-structural model incorporates regression equations using lagged values of the variable that is to be predicted. If more complicated modeling techniques cannot generate forecasts that beat these naive models, the model building attempt is presumably not worthwhile.

For people engaged in industry and finance, where having more accurate forecasts than your competitors will materially improve profitability, forecasts are useful if they provide results that are more accurate than the competition's. A model that accurately predicted the direction of change in the stock market the next day 60 percent of the time would be tremendously valuable – even though it would be wrong almost half the time – regardless of the methodology used to develop those predictions. In a similar vein, calculations by this author have shown, in some semi-annual polls of economists published in the *Wall Street Journal*, over 50 percent of the forecasts incorrectly predicted the *direction* interest rates would change over the next six months. Hence any model that could even predict the direction in which interest rates would move over the next several months would significantly improve the current status of forecasting financial markets.

Yet the decision not to forecast at all means throwing in the towel and claiming that any deviations from past trends can never be predicted. That would be the case only if the variable in question always grew at the same rate and was never subject to exogenous shocks. For even if changes are truly unexpected (an oil shock, a war, a wildcat strike, a plant explosion) forecasting models can still offer useful guidance indicating how to get back on track. Virtually everyone in a management or executive role in business or finance makes guesses about what will happen in the future. While these guesses will never be perfect, they are likely to be much improved if the practitioner combines robust statistical techniques with the ability to adjust the forecasts when actual events do diverge from predicted values.

Forecasting makes practitioners humble. That does not mean people who choose forecasting as a profession are necessarily humble; the opposite is more likely to be true. But unlike economic theories, which can often persist for decades without anyone ever being able to verify whether they are accurate or useful, forecasters generally find out quickly whether or not their opinions are correct.

Since highly visible forecasts of the overall economy or financial markets have compiled a very unimpressive track record over the past 30 years, it is sometimes argued that predicting economic variables is not a useful exercise. Indeed, most consensus forecasts of real growth, inflation, and interest rates have not been much better than from a naive model. In view of these results, some have concluded that forecasting models do not work very well.

Before reaching that conclusion, however, we should try to determine what causes these forecasting errors. For example, suppose the majority of forecasters thought interest rates would rise because inflation was about to increase. The Federal Reserve, also expecting that to happen, tightened policy enough that inflation decreased and, by the time six months had elapsed, interest rates actually fell. I am not suggesting this always occurs, but it is a reasonable hypothesis. Thus before beginning our analysis of how to reduce forecasting errors, it is useful to categorize the major sources of these errors. Some may be intractable, but others can be reduced by a variety of methods that will be explored in this book.

When the econometric model and the mechanism generating the model both coincide in an unchanging world, and when the underlying data are accurate and robust, the theory of economic forecasting is relatively well developed. In such cases, the root mean square forecasting error in the forecast period ought not to be any larger than indicated by the sample period statistics.

This does not happen very often; in the majority of forecasts, the actual error is significantly larger than expected from sample period statistics. In some cases that is because the model builder has used inappropriate statistical techniques or misspecified the model through ignorance. Most of the time, however, unexpectedly large forecasting errors are due to some combination of the following causes:

- structural shifts in parameters
- model misspecification
- missing, smoothed, preliminary, or inaccurate data
- changing expectations by economic agents
- policy shifts
- unexpected changes in exogenous variables
- incorrect assumptions about exogeneity
- error buildup in multi-period forecasts.

1.2.1 Structural Shifts in Parameters

Of the factors listed above, structural shifts in parameters are probably the most common. These may occur either within or outside the sample period. For example, sales at Ace Hardware will drop dramatically when Home Depot opens a store three blocks away. At the macroeconomic level, a recession used to be accompanied by a stronger dollar; now it is accompanied by a weaker dollar. Company profits of American Can were influenced by completely different factors after it became a financial services company.

Perhaps stated in such stark terms, structural shifts are obvious, but most of the time the changes are more subtle. For 1997 through 1999, macroeconomists thought the growth rate of the US economy would slow down from about 4% to the 2–2½% range; yet each year, real growth remained near 4%. Forecasters thought that with the economy at full employment, inflation would increase, causing higher interest rates, lower stock prices, and slower growth, yet it did not happen. At least in retrospect, there were some structural shifts in the economy. For one thing, full employment no longer produced higher inflation. Also, the technological revolution boosted capital spending and productivity growth more rapidly. Yet even after several years, the consensus forecast failed to recognize this shift.

1.2.2 Model Misspecification

Model misspecification could be due to the ignorance of the model builder; but even in the case where the best possible model has been estimated, some terms might be omitted. In many cases these might be expectational variables for which data do not exist. For example, economists agree that bond yields depend on the expected rate of inflation, a variable that cannot be measured directly. A company might find that cutting prices 5% would not invoke any competitive response, but cutting them 10% means competitors would match those lower prices. The missing variable in this case would be the trigger point at which competitors would respond – which itself is likely to change over time.

It is also possible that the underlying model is nonlinear. In one fairly straightforward and frequently documented case, purchases of capital goods with long lives (as opposed to computers and motor vehicles) generally increase faster when the rate of capacity utilization is high than when it is low. At the beginning of a business cycle upturn, capital spending for long-lived assets is often sluggish even though interest rates are low, credit is easily available, stock prices are rising rapidly, sales are booming, and profits are soaring. Once firms reach full capacity, they are more likely to increase this type of capital spending even if interest rates are higher and growth is slower.

To a certain extent this problem can be finessed by including variables that make the equation nonlinear, and I will discuss just such an example later. For example, investment might grow more rapidly when the rate of capacity utilization is above a certain level (say 85%) than when it is below that level. However, the situation is not that simple because a given level of capacity utilization will affect investment differently depending on the average age of the capital stock, so using a simple rule of thumb will generally result in model misspecification. An attempt to pinpoint the exact rate at which capital spending accelerates is likely to result in data mining and the resultant penalty of relatively large forecast errors.

1.2.3 Missing, Smoothed, Preliminary, or Inaccurate Data

The data used in estimating forecasting models generally comes from one of three major sources. Most macroeconomic data are prepared by agencies of the Federal government, including the Bureau of Economic Analysis (BEA), the Bureau of the Census, and the Federal Reserve Board of Governors. Financial market data on individual company sales and earnings are prepared by individual corporations. In an intermediate category, many industry associations and private sector institutions prepare indexes of consumer and business sentiment, and measures of economic activity for specific industries or sectors; perhaps the best known of these are the Conference Board index of consumer attitudes and the National Association of Purchasing Managers index of business conditions in the manufacturing sector.

Except for specific data based on prices given in financial markets, virtually all macroeconomic or industry data are gathered by sampling, which means only a relatively small percentage of the total transactions is measured. Even when an attempt is made to count all participants, data collection methods are sometimes incomplete. The decennial census is supposed to count every person in the US, but statisticians generally agree the reported number of people in large cities is significantly less than the actual number; many of the uncounted are assumed to be undocumented aliens. Thus even in this most comprehensive data collection effort, which is supposed to count everyone, some errors remain. It is reasonable to assume that errors from smaller samples are relatively larger.

Virtually all macroeconomic and industry data series collected and provided by the government are revised. The issuing agencies named above make an attempt to provide monthly or quarterly data as quickly as possible after the period has ended. These releases are generally known as “advance” or “preliminary” data. In general, these data are then revised over the next several months. They are then revised again using annual benchmarks; these revisions usually incorporate changing seasonal factors. Finally, they are revised again using five-year censuses of the agricultural, manufacturing, and service sectors. In addition, some of the more comprehensive series, such as GDP and the CPI, may be revised because of methodological changes.

The revisions in the data prepared and released by the Federal government are often quite large. Sometimes this is because preliminary data, which appears shortly after the time period in question has ended, are based on a relatively small sample and then revised when more comprehensive data become available. In other cases, seasonal factors shift over time. Data revisions quite properly reflect this additional information.

Most government data are collected from surveys. From time to time, respondents do not send their forms back. What is to be done? The sensible solution is to interpolate the data based on those firms that did return their forms. The problem with this approach is that, in many cases, it is precisely those firms that failed to return their forms that faced unusual circumstances, which would have substantially modified the data. Eventually the problem is solved when more complete numbers are available, but the initial data are seriously flawed.

Sometimes, the methodology is changed. In October 1999, a comprehensive data revision boosted the average growth rate of the past decade by an average of 0.4% because the Bureau of Economic Analysis (BEA) – the agency that prepares GDP and related figures – decided to include software purchased by businesses as part of investment; previously it had been treated as an intermediate good and excluded from GDP. Since software had become an increasingly important part of the overall economy, this change was appropriate and timely.

In another important example, the methodology for computing the rate of inflation was changed in the mid-1990s. As a result, the same changes in all individual components of the CPI would result in an overall inflation rate that was 0.7% lower. These changes reflected the improved quality of many consumer durables, shopping at discount malls instead of department stores, and changes in market baskets that included a higher proportion of less expensive goods. Most economists agreed these changes were warranted, and many thought they were overdue. A commission headed by former Chairman of the Council of Economic Advisers Michael Boskin calculated that the rate of inflation had been overstated by an average of 1.1% per year.[3]

The Federal government statisticians cannot reasonably be criticized for including improved information and methodology in their data releases when they become available. Indeed, failure to include these changes would be a serious error. Nonetheless, the appearance of preliminary data that are later revised substantially raises significant issues in both building and evaluating forecasting models. At least in the past, it has sometimes had a major impact on policy decisions.

For example, one of the major examples of misleading preliminary data occurred in the 1990–1 recession. During that downturn, BEA initially indicated the recession was quite mild, with a dip in real GDP of only about 2%. Subsequent revisions revealed that the drop was much more severe, about 4%.[4] Acting on the data that were originally reported, the Fed assumed the slump was not very severe and hence eased cautiously. If it had known how much real

[3] Boskin, Michael J., E. R. Dulberger, R. J. Gordon, and Z. Griliches, "The CPI Commission: Findings and Recommendations," *American Economic Review Papers and Proceedings*, 87 (1997), 78–83.

[4] This result was not entirely a surprise. Joseph Carson, an economist at Chemical Bank who had previously worked at the Commerce Department, stated at the time that he thought real GDP was declining at a 4% annual rate in late 1990.

GDP had really fallen, it probably would have eased much more quickly. Indeed, when the recovery failed to develop on schedule in 1991, the Fed did reduce short-term interest rates to unusually low levels by the end of 1992, and the economy finally did recover. However, that boosted inflationary pressures, causing the Fed to tighten so much in 1994 that real growth plunged to 1% in the first half of 1995. Not until the latter half of that year did the economic effects of those incorrect data completely disappear.

The most accurate forecast would have said the economy is in worse shape than the government reports indicate, so initially the Fed will not ease enough and hence the economy will be slow to rebound, which means the Fed will eventually have to ease more than usual, so two years from now interest rates will be much lower than anyone else expects, in which case inflationary expectations will rise and the Fed will have to tighten again. Of course no one said that, nor could anyone have reasonably been expected to predict such a sequence of events.

This example clearly indicates how inaccurate data can cause poor forecasts. Yet economists were roundly criticized for underpredicting the severity of the recession, overpredicting the initial size of the rebound, and failing to gauge the decline change in interest rates accurately. No forecaster won plaudits following that recession, but it is not unreasonable to suggest that forecast errors would have been smaller with more accurate data.

In May 1974, the wage and price controls imposed by the Nixon Administration ended. As a result, the producer price index (PPI) rose by a record amount that month. For the next few years, the seasonal adjustment program used by the Bureau of Labor Statistics (BLS) assumed the PPI always rose sharply in May, so the seasonally adjusted data for the May PPI showed a big dip, while the unseasonally adjusted data were virtually unchanged. In this case perhaps the obvious solution would have been to ignore those data, but it is not clear what method the forecaster should use. Running regression equations without any data for May? Using seasonally unadjusted data? Treating May 1974 with a dummy variable – e.g., 1 in that period and 0 elsewhere? All these are possible, but none is optimal.

Of course, it is not only the Federal government that revises their data. Companies often restate their earnings for a variety of reasons. They book sales when they ship the goods, but if they aren't paid for, writeoffs must be taken. Sometimes reorganizations, or sales of divisions, result in huge one-quarter writeoffs. Other times, accounting errors are at fault. Analysts try to take these anomalies into account, but most if not all attempts to predict stock prices based on reported company earnings suffer from the changes and inconsistencies in these data.

There will never be any perfect solution to the issue of data revisions. Nor does it make any sense to castigate government statisticians for providing the most accurate estimates possible based on incomplete data and the changing nature of the economy. Nonetheless, a few observations relating to data revisions are appropriate at this point.

1 Evaluation of forecast accuracy should take into account the data at the time when the forecasts were issued, rather than the most recently revised data. This means, for example, that an attempt to evaluate forecasting accuracy of macroeconomic forecasts made many years ago provides far different results depending on which set of "actual" data are used.
2 Some, although certainly not all, forecast error stems from the assumptions of changes in fiscal and monetary policy that are based on the preliminary data issued by the government. Later revisions of these data sometimes make it appear that those assumptions were unwarranted.
3 When estimating a structural model over an extended period of time, it is useful and appropriate to use dummy or truncated variables in the regression equation. For example, the methodological changes in the CPI that began in 1994 can be entered explicitly as an additional variable; before 1994, any such variable would have the value of zero.

1.2.4 Changing Expectations by Economic Agents

This is often cited as one of the major reasons given for the failure of macroeconomic modeling in the 1970s and the 1980s. It has been argued that economic forecasts based on past historical evidence cannot be accurate because people adjust their behavior based on previous events, and thus react differently to the same phenomena in the future. This concept is generally known as the Lucas Critique;[5] however, it was formulated by Oskar Morgenstern in 1928,[6] so it is hardly a recent idea. Formally, we can express this concept by saying that the data generation process underlying the model has changed during the sample period, or between the sample and the forecast periods. I mention the roots of this concept to emphasize that it far predates the idea that mismanaged monetary policy in the 1950s and 1960s was the primary factor that caused the short-term tradeoff between inflation and unemployment.

Indeed, the Lucas Critique is just a special case, although an extremely well-known one, of changing expectations. Economic agents often change their behavior patterns based on what has happened in the past. That is not only true at the macro level. Growth in individual company sales will be significantly affected as competitors enter and exit the industry. Firms will raise or lower prices depending on how their competitors react. Borrowers may have a higher or lower rate of default on loans depending on recent changes in the bankruptcy laws.

Lucas and others, and Morgenstern before them, claimed that econometric models would not work whenever economic agents learned from previous

[5] Lucas, Robert E., "Some International Evidence on Output–Inflation Tradeoffs," *American Economic Review*, 63 (1973), 326–34.

[6] Morgenstern, Oskar, *Wirtschaftsprognose: eine Ubtersuchung ihrer Vorussetzungen und Moglichkeiten* (Julius Springer, Vienna), 1928.

experience and adjusted their behavior accordingly in the future. Yet many economic links continue to hold over a wide variety of different experiences. On a ceteris paribus basis, consumers will spend more if income rises, although admittedly their increase in consumption will be greater if they think the change is permanent rather than temporary. If interest rates rise, capital spending will decline. If the value of the currency increases, the volume of net exports will decline. If the growth rate for profits of an individual firm accelerates, the stock price will rise. There are many similar examples where structural relationships continue to hold in spite of changing expectations.

Sometimes, a change in expectations in one area of the economy will generate changes in other sectors that are consistent with past experience. One major example of this occurred in the US economy in the second half of the 1990s. Expectations about future profit growth shifted significantly, so that the price/earnings ratio of the stock market doubled even though bond yields were at just about the same level in 1995 and 2000. Few forecasters were able to predict that change. On the other hand, the rise in stock prices and the decline in the cost of equity capital impacted consumer and capital spending in a manner consistent with previous historical experience. In addition, the more rapid growth in capital stock stemming from an increase in the ratio of capital spending to GDP boosted productivity growth, which reduced the rate of inflation and lowered interest rates further. That in turn boosted real growth enough that the Federal budget position moved from a deficit to a surplus, which further boosted equity prices. Predicting the change in the stock market was difficult; but given that change, predicting more robust growth in the overall economy was more straightforward. Conversely, when the stock market plunged, all of the reverse factors occurred – lower capital spending, a slowdown in productivity, and a return to deficit financing.

1.2.5 Policy Shifts

Anyone who tries to estimate an equation to predict short-term interest rates will soon find that, during the mid-1970s, Fed Chairman Arthur Burns used monetary policy to offset the recessionary impact of higher oil prices, leading to unusually low real interest rates; whereas in the early 1980s, Chairman Paul Volcker refused to accommodate the further increase in oil prices, leading to unusually high real interest rates. The real Federal funds rate equals the nominal rate minus the change in inflation over the past year. Its pattern is shown in figure 1.1.

No model estimated on data through 1979 would have predicted the massive increase in real interest rates that started in late 1980. With hindsight, of course, one can include a well-chosen set of economic variables that track this pattern, but that is not the point. In July 1980, the Blue Chip consensus forecast of the six-month commercial paper rate for 1981 was 8.7%; the actual figure was 14.8%. This is one of the clearest policy shifts that ever occurred in the US economy.

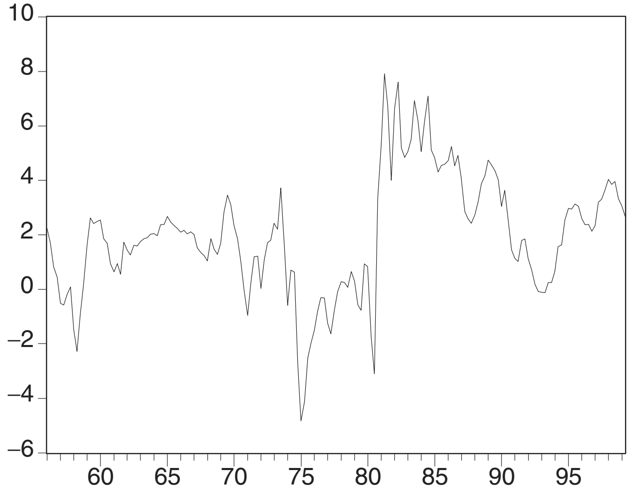

Figure 1.1 The real Federal funds rate.

What lessons can forecasters learn from this experience? In the short run, fluctuations in short-term interest rates are determined primarily if not exclusively by the action of the Federal Open Market Committee. That is why short-term interest rate forecasting today is reduced to a series of guesses about what the Fed will do next. In the long run, however, we learn another lesson. If the Fed holds short-term rates at below equilibrium for an extended period, eventually both inflation and interest rates will rise; whereas if it holds short-term rates above equilibrium, eventually both inflation and interest rates will decline. In this case, a model that captured this underlying relationship would provide very little guidance in predicting interest rates in the short run, but would be useful in the long run. In particular, a forecast that interest rates and inflation would start to decline in 1982, hence setting in motion the biggest bull market in history, would have been particularly valuable. Yet hardly anyone believed the few forecasters who accurately predicted that development.

Even the best econometric model is not designed to predict the impact of unexpected policy or exogenous changes in the short run. However, once these changes have occurred, correctly structured models should be able to offer valuable insights into what will happen in the longer run.

1.2.6 Unexpected Changes in Exogenous Variables

The change in Fed policy under Paul Volcker is a classic example of a policy change initiated by the government. As seen by forecasts made at the time, it was a major surprise. Another major example of an unexpected policy change, although it occurred over several years, was the decision by senior officials in the Clinton Administration to reduce the level of real per capita government spending during his tenure as President.[7] That changed the budget deficit to a budget surplus, which (as already noted above) was one of the factors causing an almost unprecedented increase in the price/earnings ratio of the stock market.

Changes of this sort are undertaken by government officials. However, other shocks that affect the economy have nothing to do with policy, such as energy shocks, wars, and natural disasters. Unless foreseen, they will not be incorporated in any forecasts. Yet if they were predicted, vigorous action would be taken to offset or eliminate these developments.

I have already noted how the Fed acted quite differently to the first and second energy shocks in 1973–4 and 1979–80 respectively. However, that was not the only change; private sector economic agents also reacted differently. The first energy shock was viewed by most consumers and businesses as a once-in-a-lifetime event, so they did not alter their behavior patterns very much. As a result, oil imports continued to increase, and eventually oil prices rose again. After the second energy shock, attitudes changed significantly. Most people now expected that massive price increases would continue on a regular basis, and forecasts were common that oil prices would rise to $100/bbl by the end of the twentieth century. As a result, both consumers and businesses started using less energy, buying more fuel-efficient motor vehicles, and constructing more fuel-efficient buildings. Those plans were successful enough to reduce oil imports, so in 1986 energy prices plunged by more than half. In 1998 they were lower in real terms than in 1972, before the first energy shock occurred.

Any forecast of the economy in the 1980s – whether right or wrong – was influenced by the assumption about energy prices. However, this example indicates the value of some of the alternative types of forecasts discussed in section 1.3: conditional vs unconditional, point vs interval, and alternative scenarios weighted by probabilities. An appropriate way for many businesses to proceed would have been to generate alternative forecasts based on different scenarios about oil prices: higher, steady, or lower. When prices gradually started to decline in the mid-1980s as the worldwide energy glut increased, more weight would have been given to the lower-price scenario, so businesses would have been better prepared when crude oil prices suddenly fell by more than half in 1986.

There is little to be gained by pointing out that forecasts are inaccurate when they fail to predict unexpected exogenous shocks, many of which would never have occurred if they had been accurately predicted. However, models that correctly analyze the impact of these shocks when they do happen can still be quite useful in indicating what lies ahead.

---

[7] Most of these changes were suggested by Treasury Undersecretary, and then Secretary, Robert Rubin.

1.2.7 Incorrect Assumptions About Exogeneity

In some cases, models designed for forecasting generate much larger errors than would be indicated by the sample period statistics because some of the independent variables are assumed to be exogenous when they really are not. Technically, an exogenous variable is one whose value is not determined within an economic model, but which plays a role in the determination of the endogenous variables. However, as a practical matter, there are degrees of exogeneity. Only a relative handful of variables, such as weather conditions and defense expenditures, are exogenous in all circumstances. Most of the time, policy variables have some endogenous components as well.

For example, foreign currency values are often considered to be exogenous. After the collapse of the Thai baht, Korean won, Indonesian rupiah, and Malaysian ringgit in the latter half of 1997, US net exports declined dramatically in 1998 and the first half of 1999. As a result, manufacturing production rose much more slowly than total GDP; whereas during boom years, production usually rises faster than overall GDP. North Carolina has the highest proportion of workers in manufacturing, so its growth rate fell sharply after the collapse of those currencies.

A model that linked growth in North Carolina employment to the value of the dollar (among other variables) would show a high correlation. However, a forecast made in 1997 would have been inaccurate if it had assumed the values of those currencies would remain stable. In such a case, the model would appear to work well, but forecasts of the North Carolina economy would be far off the mark. In this case, the equations might have continued to work well in the sense of high correlations and low standard errors, but the forecasts would have been poor because of the inability to predict the exogenous variables.

In the past, monetary policy used to be treated as exogenous, although this error is made far less often today. Even in the days before Paul Volcker, the Fed routinely tightened monetary policy as inflation increased. Thus assuming that monetary policy variables were exogenous and would not change invariably led to forecast errors that were much larger than expected from the sample period statistics.

1.2.8 Error Buildup in Multi-period Forecasts

Analyses of macroeconomic models undertaken many years ago by this author showed that the single biggest source of error in multi-period forecasting was caused by using the lagged dependent variable on the right-hand side of the equation. If current consumption were estimated as a function of lagged consumption, for example, an error made one quarter could distort all the forecasts from that point forward. I will discuss a variety of methods to overcome that difficulty; now that this error has been well documented, it does not occur so much in multi-period forecasting models. Nonetheless, it is an error that beginning modelers often commit.

1.3 Alternative Types of Forecasts

When most people think of forecasts, they think of point estimates. For example, sales will rise 12% next year, the Dow will climb to 12,000 a year from now, the Federal Open Market Committee will vote to boost the Federal funds rate 25 basis points at its next meeting, the price of oil will climb 20% over the next six months, and so on.

While it is true that point estimates are the most common type of forecasts, there are many other ways in which forecast results can be presented. Sometimes a range for the predicted variable is more appropriate; other times, different probabilities are assigned to alternative scenarios. Sometimes the penalties associated with being too high or too low are equal; at other times, the loss function is asymmetric. In this section, I discuss some of the more common types of alternative forecasts.

1.3.1 Point or Interval

Suppose a company has a limited amount of excess manufacturing capacity. If sales grow less than 5% per year, the company will be better off using its existing facilities. If sales grow more than 5% per year, it will be better off building a new plant. In this case, the point estimate for sales growth is not as important as the probability that sales growth will exceed 5%.

A similar case might be made for advertising budgets. If a firm thinks a \$1 million expenditure on advertising will boost sales by at least \$5 million, it will decide to go ahead and spend the money. It doesn't matter so much whether the increase in sales is \$6 or \$10 million, but if it is \$4 million, the expenditure will not be made.

At the macro level, suppose the Fed decides that 3% is the highest level of inflation that is tolerable. If inflation rises 1%, 1½%, or 2%, there will be no change in monetary policy. If it exceeds 3% – or if it appears likely it will soon exceed 3% if policy is not changed – the Fed will boost short-term interest rates.

A company may have a loan covenant with the bank stating that if cash reserves drop below a certain level, the loan will be called. That level might be correlated with the assumption of increased profitability, so a decline in profits would trigger the loan call. In that case, the key forecast is whether company profits have risen or not, rather than the precise amount they would increase.

1.3.2 Absolute or Conditional

Forecasts can be either absolute or conditional. Some examples of absolute, or unconditional forecasts are: real GDP will grow 4% next year, the Republicans will retain (or regain) majority control of Congress, and company sales will rise at least 15% per year over the next decade. However, many forecasts are issued on a conditional basis: real GDP will grow 4% next year if the Fed does not tighten, the Republicans will be the majority party in Congress if they also capture the Presidency, and sales will grow if competitors do not double their capital spending and advertising budgets.

The choice of which type of forecast is appropriate will depend largely on how the results are to be used. A speculator in financial markets wants to know whether prices will rise or fall, not whether they will rise or fall under certain circumstances. An automobile dealer wants to know what lines of vehicles will sell most quickly, so he can optimize his ordering procedure. A pharmaceutical company wants to know how rapidly a new drug will be adopted.

Conversely, conditional forecasts can often be quite useful. Firms might want to determine how fast sales are likely to grow under normal business conditions, using those results as guidelines for rewarding superior performance. If sales are then affected by some exogenous event, guidelines can be adjusted accordingly. Forecasts of production planning might be determined based on the assumption that materials are delivered on time, compared with what might happen if a strike occurred. The most common way of delivering conditional forecasts is by using alternative scenarios, as discussed next.

1.3.3 Alternative Scenarios Weighed by Probabilities

A forecast that sales will rise 8% if the economy booms, rise 6% if real growth remains sluggish, and fall 2% if there is a recession may appear to be an excuse to avoid offering a firm forecast at all. However, that is not always true. In many cases, firms need to be prepared to take appropriate action if the economy falters even if the probability of that occurring is relatively low.

Based on the historical forecasting record of macroeconomists, it would appear that recessions were not predictable. Consider the case of a lending institution involved in sub-prime auto loans. As long as the economy remains healthy, the vast majority of these loans will be repaid; if a recession strikes, the loss rate will rise enough to put the company out of business. Prudence might dictate less risky loans; but if the company is too picky, it will lose business to competitors and won't make enough loans to stay in business.

In this case the most appropriate procedure would be to assess the probability of a recession occurring next year. If it were only 5%, then the lending institution would continue to expand its sub-prime loan portfolio. On the other hand, if it were to rise to 25%, some trimming would be in order. Note that in this case the probability of an actual downturn the following year is well below 50%, yet some adjustment in corporate strategy is warranted.

The alternative-scenario method of forecasting can also be used for long-range planning, since long-term economic forecasts are generally little more than trend extrapolations in any case. The company might discover that the probability of meeting its stated goal of a 15% annual gain in sales and earnings would occur only if the most optimistic macroeconomic forecast, with a probability of only 10%, were met. The company could then make plans to move into faster-growing areas of the economy or, alternatively, trim its ambitious long-term goals to more realistic levels.

1.3.4 Asymmetric Gains and Losses

So far we have been assuming that a forecast error of +8% carries the same penalty as an error of –8%. Often, however, that is not the case. For many companies, if sales increase faster than expected, that is fine; but if they don't, disaster strikes. I have already described such a situation for a sub-prime auto lending company. The same general type of argument could be applied to municipal bonds; as long as the community tax base grows above a certain rate, the interest and principal will be repaid, but if it dips below that rate, the issuing authority will default on the bonds.

In many companies, the rewards for exceeding the plan are substantial: bonuses, promotions, and larger budgets for next year. Similarly, the penalties for failing to meet planned targets are severe, including loss of employment. In a situation of that sort, many planners will set targets below their predicted level, so they will appear to have exceeded their goals. Eventually, management may catch on to this trick and fire all the planners, which is another risk. Nonetheless, the percentage of plans that are exceeded compared with the percentage that are not met strongly suggests that corporate planners are well aware of the asymmetric loss function.

Money managers may face a similar dilemma. If they beat the benchmark averages – Dow Jones Industrials, S&P 500, or Nasdaq composite index – they are handsomely rewarded; investors will switch their assets into those funds, and salaries and bonuses rise. If their performance falls short of the gains posted by the major averages, they will lose customers and possibly their own jobs.

This is not just a hypothetical example. The so-called January effect occurs because many money managers aggressively buy growth stocks early in the year (or the previous December) and, if they can show substantial gains, lock in those gains and buy the equivalent of index funds for the rest of the year. In the same vein, very few money managers who are already ahead of the average for the first three quarters of the year would take risks in the fourth quarter that would jeopardize their hefty year-end bonuses.

1.3.5 Single-period or Multi-period

So far we have not specified how many time periods in the future are being predicted. That can make a great deal of difference in the way a model is formulated. In models used to forecast only one period ahead, it might well be appropriate to use the lagged value of the variable that is being predicted. Interest rates in the next period might very well depend on rates this period, as well as on other variables such as the inflation rate, growth rate, unemployment rate, value of the currency, budget surplus or deficit, and other relevant variables.

However, suppose the model is used to predict interest rates on a monthly basis for the next 12 months. In this case, the forecasts for interest rates later in the year would depend on "lagged" values of interest rates that were not known at the time of forecast. For example, suppose the forecast made at the beginning of March for interest rates depends on the level of interest rates in January and February. As the year progresses, the forecast for interest rates in June would depend on their level in April and May, which are not yet known.

For this reason, using the lagged dependent variable for multi-period forecasts causes serious difficulties that do not exist for the single-period forecast. That does not rule out the use of lagged dependent variables on an a-priori basis, but it does raise a red flag. One of the tenets of the classical linear model, as will be shown in the next chapter, is that the values of all the independent variables are known at the time of forecast. Obviously that is not the case when the lagged dependent variable is used in multi-period forecasting. Hence it is advisable to use a different approach when multi-period forecasts are required.

1.3.6 Short Run or Long Range

To a certain extent, the difference between short- and long-run forecasts can be viewed as the difference between single- and multi-period forecasting. However, whereas short-term forecasts are more generally concerned with deviations from trends, long-run forecasts are often designed to predict the trend itself. As a result, different methods should be used.

One of the principal goals of short-term forecasting, and one that has been emphasized by time-series analysis, is to remove the trend from time-series variables so the underlying properties of the series may be properly examined. If company sales have been growing an average of 12% per year, the challenge in short-term forecasting is to indicate how much sales next year will deviate from that trend. Long-range forecasters, on the other hand, might want to determine how many years it will take for the trend growth in sales to diverge from that 12% average gain. The difference is analogous to the split responsibilities of the COO, who asks "How are we doing?", and the CEO, who asks "Where are we heading?"

In large part, then, the method of building forecasting models will be different depending on whether the primary goal is short-term or long-range forecasting. In general, the same model will not be optimal for attempting both goals.

1.3.7 Forecasting Single or Multiple Variables

In the models discussed above, it has been implicitly assumed that the independent variables – the variables on the right-hand side of the equation – are either known in advance or are truly exogenous. In the case of financial decision or qualitative choice models, actual information is entered for economic and demographic data. In the case of sales forecasting models, the variables are either exogenous to the firm or are determined by management decisions.

In the case of macroeconomic and financial forecasting models, however, that assumption is not generally valid. Interest rates depend on expected inflation, which is generally not known. Net exports depend on the value of the currency, which also is not known. In cases of this sort, it is necessary to build multi-equation models in order to explain all the endogenous variables in the system. In the case of macro models, some variables are generally treated as exogenous, such as changes in fiscal and monetary policy, but even these are often related to the state of the economy. Only variables such as wartime expenditures, energy shocks, or weather conditions are truly exogenous.

1.4 Some Common Pitfalls in Building Forecasting Equations

Before turning to a brief review of statistics, I will illustrate some of the most common pitfalls that occur in estimating regression models for forecasting. These topics will be treated in a more rigorous fashion after the statistical groundwork has been prepared, but it is useful to introduce them initially so they can be kept in mind as the statistical and econometric exposition unfolds.

I have already noted that there is no such thing as a perfect forecast. Even if all of the statistical methods are applied correctly, some random error will occur. This error can be quantified and measured for any existing data set, and can be used as an estimate of the forecast error that can be expected. In the vast majority of cases, though, the actual forecasting error is larger than is indicated by the regression equation or econometric model. Some of the major reasons for unexpectedly large forecast error are discussed next.

The residuals in any stochastic equation, which are supposed to be independent, may be correlated with each other. As a result, there are far fewer independent observations than indicated by the statistical program. Hence the goodness-of-fit statistics are overstated, and the forecasting errors are understated. Structural relationships estimated with time-series data – consumption as a function of income, prices as a function of unit labor costs, or interest rates as a function of the inflation – are all likely to have serially correlated residuals. Because consumer spending patterns, for example, change slowly over time, the number of independent observations is probably far less than the sample period data would indicate. Consequently, the standard errors are significantly understated.

Virtually all statistical and econometric tests are based on the underlying assumption that the residuals are normally distributed. Often, however, that is not the case. That is another reason why the calculated goodness-of-fit statistics overstate the robustness of the equation.

The "law of large numbers" indicates that as the sample size increases, all distributions with a finite variance tend to approach the normal distribution. However, that is scant comfort to those who must deal with relatively small samples. Furthermore, some financial market data do not have bounded data; in particular, percentage changes in daily stock prices are not normally distributed. Every once in a while, an unexpected event will cause a much larger change than could be expected from past history – especially in financial markets. Such distributions are colloquially referred to as "fat tails." Estimates based on the assumption of a normal distribution when that is not the case are likely to generate disappointing forecasts.

Spurious correlation may destroy the usefulness of any model for forecasting, even if the sample period statistics appear to provide a remarkably accurate fit. Many studies have shown that series that actually have no correlation – because they were generated from random number series – can provide highly significant goodness-of-fit statistics if enough alternative regressions are calculated. This problem has become particularly virulent in the PC era, where it is a simple matter to run hundreds if not thousands of regression equations very quickly.

The problem of "data mining" has also run rampant because of quick and inexpensive computing power. This issue always represents somewhat of a dilemma. One does not want to test only one or two versions of any given equation. After all, the theory may not be precisely specified; and even if the long-run determinants are well determined, the lag structure and adjustment process may not be known. Empirical approximations of theoretical concepts may not be precise, so it is logical to try several different measures of the concept in question. Also, research results are often improved when alternative specifications were tried because the first attempt did not produce reasonable results. Yet having provided all these reasons for diligent research, it is much more likely that econometricians and statisticians will "torture the data until they confess" instead of failing to calculate the necessary minimum number of regressions. Such attempts at curve fitting seldom produce useful forecasting equations.

Sometimes the equation fits very well during the sample period, and the goodness-of-fit statistics hold even in the forecast period, yet the equation generates very poor forecasts because the values of the independent variables are not known. For example, sales growth for a particular company or individual product line is likely to change if competitors react to an erosion of their market share. At the macroeconomic level, financial markets certainly will react differently to anticipated and unanticipated changes in policy. Consumers are likely to alter their spending patterns based on what they think will happen in the future as well as changes in current and lagged income and monetary conditions.

It is not very helpful to develop theories that produce optimal forecasts under severely stylized sets of assumptions that are rarely encountered in the real world. Practical business forecasting invariably consists of two interrelated steps: use of standard statistical theory that has been developed based on restrictive assumptions, followed by modification of that theory to improve actual forecasting accuracy. These two steps cannot be considered in isolation. Thus even in this introductory chapter, I have pointed out some of the major pitfalls that commonly occur in forecasting models. Further details will be provided throughout the text.

The following examples are indicative of many cases where robust economic theories, which have been verified by sophisticated econometric methods, do not generate accurate forecasts unless they are further modified.

- *Example 1.* Economic theory says that the riskless long-term interest rate is related to the underlying growth rate of the economy, the Federal budget deficit ratio, and the expected rate of inflation. Econometrics can be used to test this theory. However, it cannot be used for forecasting unless, in addition, we can find an accurate way to predict the expected rate of inflation. Essentially the same comments could be made for forecasting the stock market, foreign exchange rates, or commodity prices. Since inflationary expectations are not formed in a vacuum, they could presumably be tied to changes in economic and political variables that have already occurred. So far, no one has been very successful at this.
- *Example 2.* The price of oil is tied to the world demand and supply for oil, which can perhaps be predicted accurately by econometric methods, using the geopolitical situation of Saudi Arabia *vis-à-vis* the US and other major powers as a major factor in the forecast. However, world economic hegemony cannot be predicted econometrically – and probably cannot be predicted very well with any method – so this is not a useful forecasting model. Certainly no one in the early 1980s publicly predicted the fall of the Berlin Wall by the end of the decade.
- *Example 3.* Historically, the growth rate for PCs, modems, and other high-tech equipment can be accurately tracked over the sample period by identifying the time when major innovations were introduced and matching their performance to various growth curves. In the future, since the timing of such innovations is unknown, such a set of regression equations would not serve as a useful forecasting model.
- *Example 4.* Economic theory says that the value of the dollar depends on relative real interest rate differentials; the *higher* the real rate in the US, the more likely it is that the dollar will appreciate. However, economic theory also says

that a stronger dollar will attract capital from abroad, hence resulting in a *lower* interest rate than would otherwise occur. Both of these theories can be verified separately, but unless further adjustments are made they are useless for predicting either the value of the dollar or interest rates, since they lead to opposite conclusions. This is indicative of a larger problem in forecasting, where an individual theory may provide robust empirical results *in isolation* but may be useless for forecasting because the factors that are being held constant in the theory are in fact always changing.

These examples provide a flavor of the problems of building a practical forecasting model. Many of the examples involve interrelationships between several variables that must be predicted simultaneously. However, even in the cases where the independent variables are actually known ahead of time, and in that sense are truly exogenous, model builders often go astray by failing to realize the spurious correlation introduced by common trends in several of the time series.

Using econometrics to build forecasting models is deceptively difficult. As Clive Granger has put it, "econometric modeling is an activity that should not be attempted for the first time."[8] It takes practice to develop useful forecasting models.

Problems and Questions

1. As an economist, you are asked to prepare quarterly forecasts for the next two years for shipments of oil-drilling equipment. Data on company and industry shipments are available back to 1959. Figure 1.2 shows the relationship between constant-dollar shipments of oil-drilling equipment and the relative price of crude oil.

(a) Would you prepare an unconditional or conditional forecast? If the latter, for what variables would you prepare alternative scenarios?
(b) How would you generate forecasts of oil prices?
(c) In general, would you predict that the next time oil prices rise sharply, shipments of oil-drilling equipment would rise rapidly as they did in the 1970s and the 1980s?

2. The loan portfolio of a bank has been growing at an average of 10% per year. The bank officers would like to expand growth to 15% per year,

continued

[8] Granger, C.W. J., *Forecasting in Business and Economics*, 2nd edn (Academic Press, San Diego, CA), 1989.

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:

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

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 $\overline{X}$, where

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

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 $\mathrm{E}(\overline{X}) = \mu_X$.

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

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

[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 $\overline{X}$ was an unbiased estimator of the expected value, it turns out that the sample variance $(X - \overline{X})^2$ 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

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

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

$$\mathrm{Cov}(X,Y) = \mathrm{E}(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))]. \tag{2.5}$$

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

$$\mathrm{Cov}(X,Y) = \sum \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N-1}. \tag{2.6}$$

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

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

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

$$Y_t = \alpha + \beta X_t. \tag{2.8}$$

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).

Degrees of freedom	*t-ratio for 5% significance*
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

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

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

$$t^2 = F. \tag{2.10}$$

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 easy 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. $\overline{R}^2$ is equally suitable for multiple regression equations, as will be seen in chapter 3. It is defined as

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

This is similar to, but not exactly the same as

$$R^2 = \frac{\text{explained variance}}{\text{total variance}}. \tag{2.12}$$

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 $\overline{R}^2 = 1.00 - 0.05 * (49/48)$, which is 0.949.

When *n* is large, R^2 is large, and *k* is small, there is very little difference between $\overline{R}^2$ and R^2. However, as R^2 drops, the difference can be substantial, especially for small samples. In extreme cases, $\overline{R}^2$ 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.

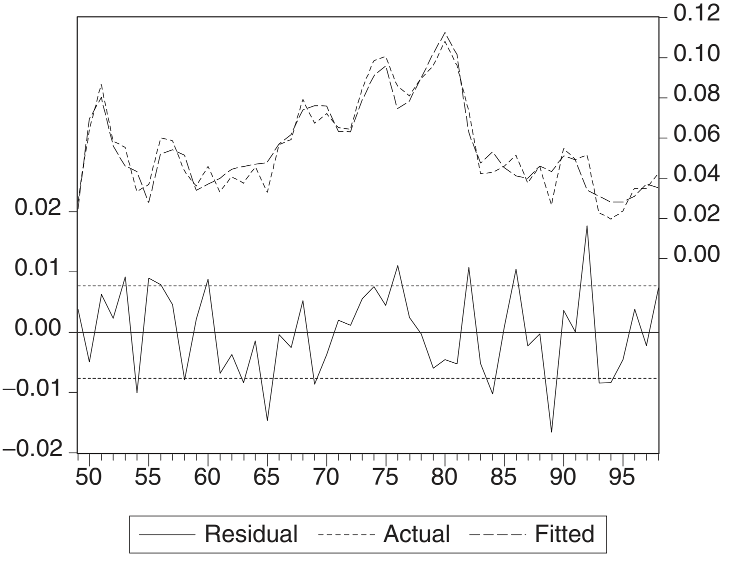

Figure 2.2 A typical output from EViews. See the text.

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

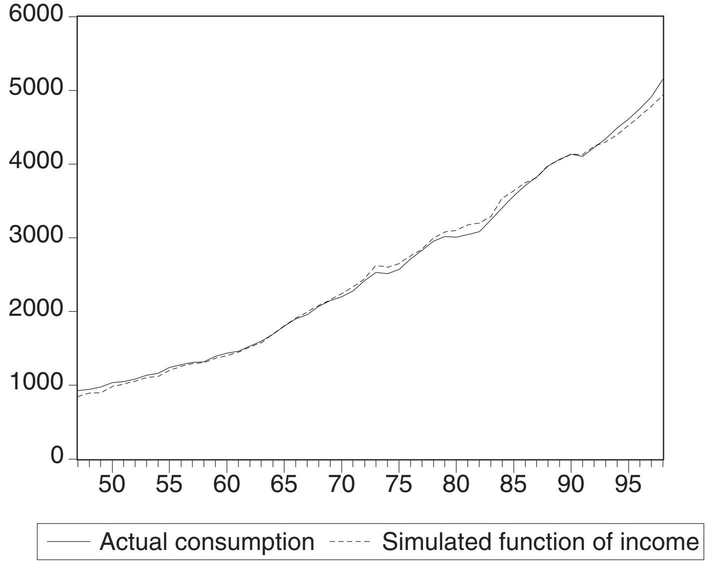

Figure 2.3 The level of real consumer spending appears to follow real disposal income very closely.

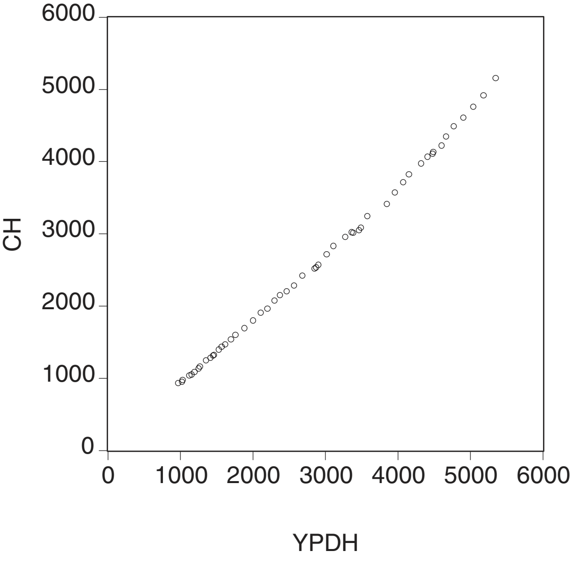

Figure 2.4 The scatter diagram between consumption and income shows almost a perfect fit.

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.

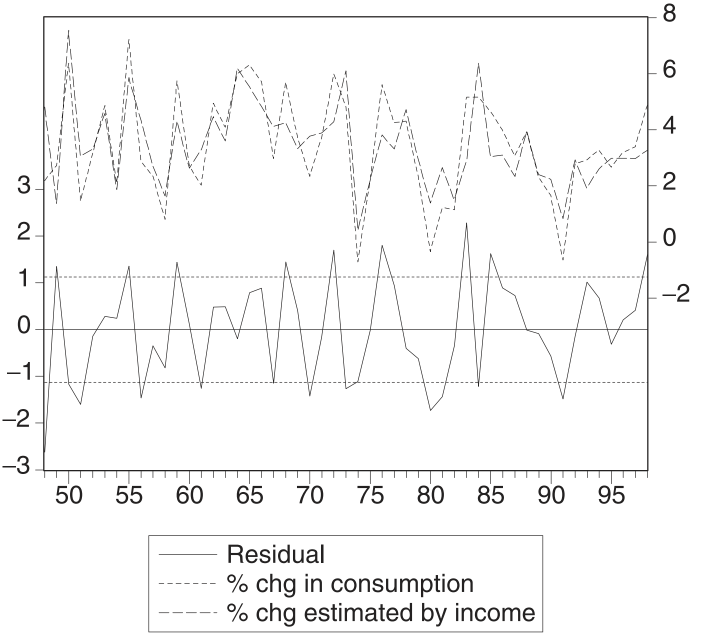

Figure 2.6 The percentage change in income is an important determinant of the percentage change in consumption, but the residuals are still quite large.

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.

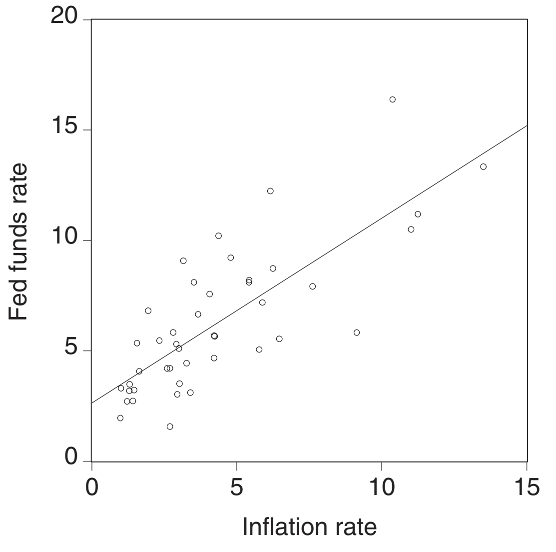

Figure 2.7 When inflation rises, the Fed funds rate also increases, but not quite as rapidly.

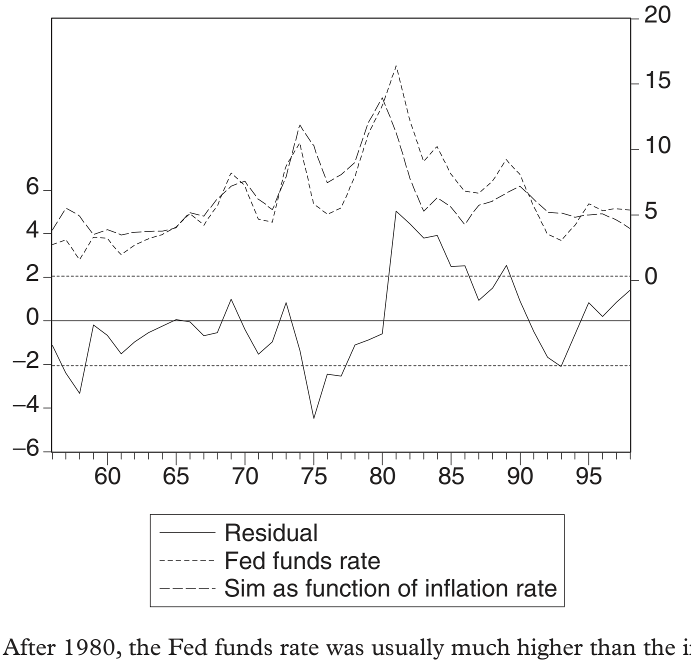

Figure 2.8 After 1980, the Fed funds rate was usually much higher than the inflation rate.

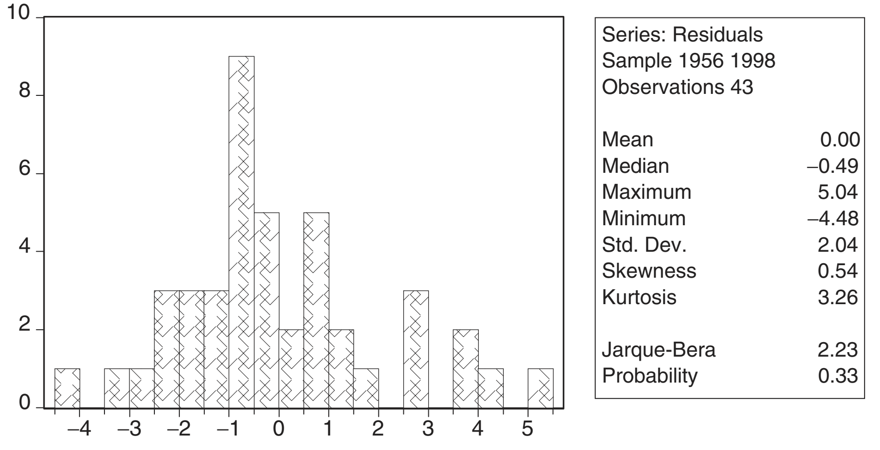

Figure 2.9 When annual data are used, the residuals from the equation where the Fed funds rate is a function of the inflation rate are normally distributed.

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

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

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

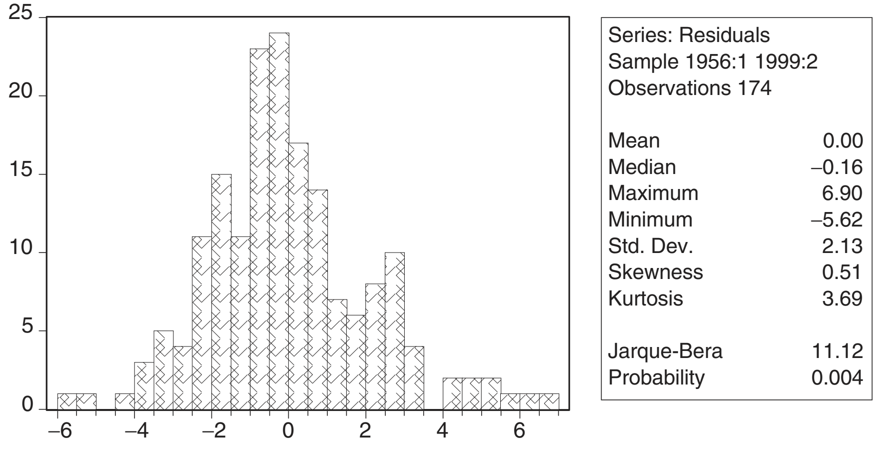

Figure 2.10 When quarterly data are used, the residuals from the equation where the Fed rate is a function of the inflation rate are not normally distributed.

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.

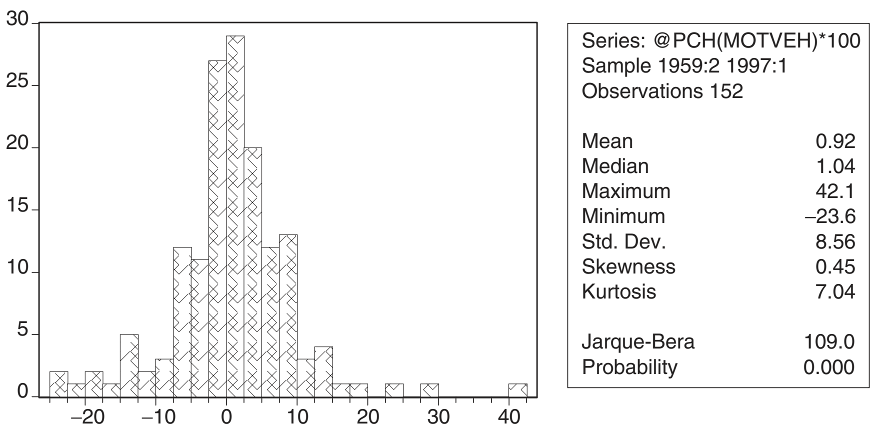

Figure 2.11 Histogram of percentage changes in quarterly motor vehicle sales.

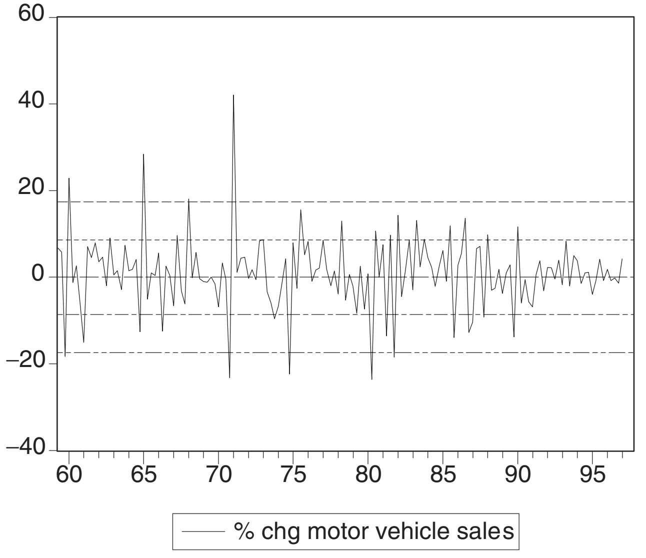

Figure 2.12 Percentage change in quarterly motor vehicle sales. The horizontal lines are 2, 1, 0, –1, and –2 standard deviations from the mean.

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

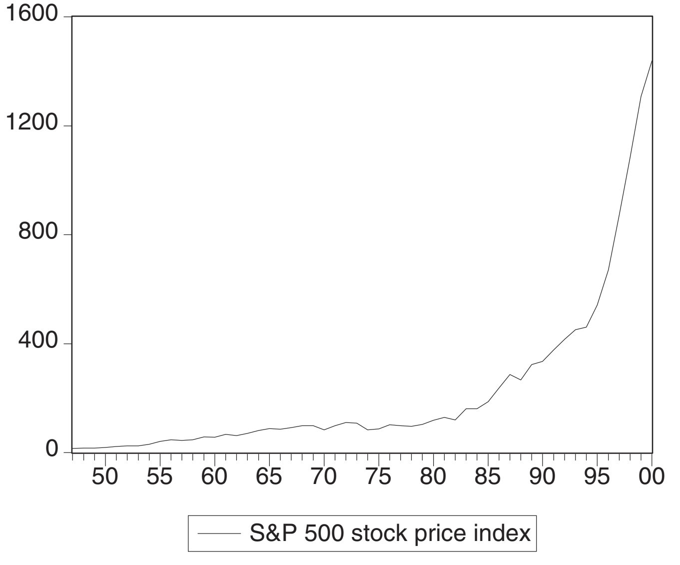

Figure 2.13 Using levels, the S&P 500 stock price index seems to be increasing at an ever-faster rate.

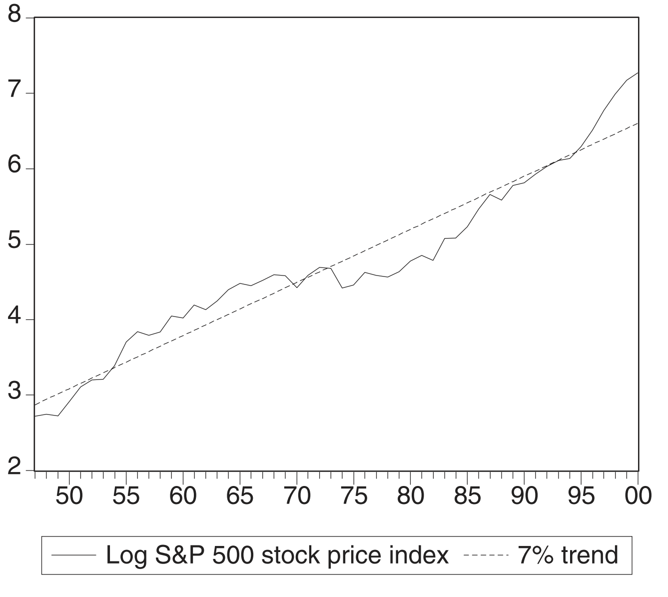

Figure 2.14 The logarithmic version of figure 2.13.

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:

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

can be written as

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

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

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

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

CHAPTER 3

THE GENERAL LINEAR REGRESSION MODEL

INTRODUCTION

One common tendency of many model builders is to calculate a large number of regression equations for a given dependent variable and then choose the equation with the highest $\bar{R}^2$. That is often a suboptimal procedure, for several reasons.

In the first place, changing the form of the dependent variable (e.g., from level to percentage change) may reduce R^2 but may also reduce the standard error. It is important not to compare apples and bicycles. Sometimes an equation with a lower R^2 will provide much more accurate forecasts.

In many forecasting models, the residuals are not normally distributed. When that happens, the goodness-of-fit statistics are invariably overstated. In particular, many models built with time-series data have residuals that are autocorrelated. When that happens, the t-ratios and R^2 statistics are overstated, so forecast errors tend to be much larger than indicated by the equation.

First-order autocorrelation of the residuals is tested using the Durbin–Watson (DW) statistic. *DW* can range from 0 to 4; if it is near 2, no autocorrelation exists; if it is below 1.4, the residuals are positively autocorrelated. A standard method exists for adjusting for autocorrelation, known as the Cochrane–Orcutt transformation. However, an equation transformed in that manner often generates inferior predictions, especially for multi-period forecasting. If *DW* is very low, this transformation is almost the same as using a first-difference equation – but the R^2 statistic is based on the levels form of the dependent variable, thus creating a highly unrealistic estimate of how accurate the forecasts will be.

The residuals may also exhibit heteroscedasticity, which often means they are dominated by a few extreme outliers. If heteroscedasticity of the residuals is present, the goodness-of-fit statistics are overstated, but that is not the major problem. Least-squares regressions give the highest weight to extreme observations, which often distorts the coefficients during normal times. Hence such an equation would be suboptimal for forecasting unless the same unusual conditions that created the outlying observations were repeated in the forecast period.

3.1 The General Linear Model

The general linear model outlined below is based on the use of least squares to estimate the parameters, their level of significance, and the goodness of fit of the overall equation. The bivariate case is presented first, followed by a discussion of some of the desirable qualities of the estimated parameters, before moving to the more general case with several independent variables.

3.1.1 The Bivariate Case

The bivariate linear model can be written as

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{3.1}$$

where Y is an observable random variable, X is a fixed or non-stochastic variable (i.e., known at time i), α is the constant term in the equation (to be estimated), β is the slope of the line relating Y and X (to be estimated), ε is a random error term with mean 0 and variance σ^2, and all the ε_i and ε_j are uncorrelated. The i subscript means there are i observations, i = 1 to T. (For example, in a time series, that might be 1955.1 through 1999.4.)

We need to explore the concept of X being a "non-stochastic" variable in this context. By definition, it is uncorrelated with the error term ε. But what does that mean?

- X could be a strictly exogenous variable, such as a time trend or a dummy variable.
- X could be a lagged variable. Since it occurred in some previous time period before Y, the error component of X would not be correlated with the error component of Y. To look at this another way, changes in X could influence changes in Y – indeed, that is what we expect to find by using it in the regression equation – but changes in Y could not influence changes in X, because X already happened.
- X could be an exogenous variable in the economic sense, such as defense spending. While the level of defense spending may indeed influence the level of Y (say GDP), the value of Y at any time will not determine defense spending, which is tied to world political considerations. Politicians do not vote to increase defense spending because the economy is in a recession and needs to be stimulated – unless they are planning to start another world war!

In most standard time-series equations, though, this assumption about X is *not true*. For example, income influences consumption, but consumption also influences income because if consumers boost their spending, more output will be produced, hence raising income. Also, current income is not known at the time of forecast. Stock prices are positively correlated with bond prices, but a drop in stock prices may cause a "flight to quality," hence boosting bond prices. The price of gasoline influences the consumption of gasoline, but if consumption of gasoline rises enough, OPEC may decide to boost its prices.

The problem of two-way correlation, which is sometimes known as simultaneity bias and sometimes as the identification problem, is a serious one in building forecasting models and is discussed later in this book. First, however, the theory and operating rules are developed under the simpler assumption that Y does not influence X.

3.1.2 Desirable Properties of Estimators

The previous chapter provided unbiased estimates of the mean and variance, μ and σ^2. It is desirable for these estimates to be consistent and efficient as well.

Unbiasedness means that the expected value of the variable is equal to the population mean. That is certainly one desirable quality of statistical estimators, or parameter estimates. However, it is not the only one.

Consistency means that the error diminishes as the sample size increases. One would certainly expect that a sample size of 100 would have a smaller standard error than a sample size of 10. This term is similar in most cases to asymptotic unbiasedness, which means the bias falls to zero as the size of the sample increases. There are a few odd probability distributions where the two terms are not the same, but for our purposes they may be considered equivalent.

While presumably no researcher wants a "biased" estimate, consistency is actually more important to statisticians than bias. Small sample sizes (often in the range of less than 20 observations) generally do not give robust estimates anyhow, but it is critical that as the sample size grows, the error diminishes. Otherwise there is no reason to suppose that the researcher is zeroing in on the correct value.

Efficiency is the other important criterion; that means the estimate in question has a smaller variance than any other estimate for a given sample size.

Sometimes efficiency is more important than unbiasedness. Consider the case where a mugger has attacked you and there are two witnesses. The actual height of the mugger is 5 ft 10 in. The first witness thinks the mugger was 5 ft 10 in, but isn't sure; his height could have been anywhere between 5 ft 2 in and 6 ft 6 in. That is an unbiased estimate, but not very useful. The other says his height was between 5 ft 8 in and 5 ft 9 in, whereas in fact it turns out to be 5 ft 10 in. That is a biased estimate but more useful.

In the same vein, a biased forecast that the stock market would rise 6% to 8% next year when in fact it rose 10% would be much more useful than an unbiased forecast that the change would be between –10% and +30%.

3.1.3 Expanding to the Multivariate Case

The general regression model can be written in a form analogous to the simple bivariate regression model, as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + \varepsilon_i \tag{3.2}$$

where Y is the dependent variable, the X_k are the independent variables, β_1 is the constant term, or intercept, of the equation, the other β are the parameter estimates for each of the X terms, ε_i is the error term, and there are i observations ($i = 1$ to T).

The assumptions for the general linear model are as follows.

- The underlying equation is linear.
- The X_i are non-stochastic, which means they are uncorrelated with the error term.
- The error term is normally distributed.
- The error term has constant variance (e.g., it does not increase as the size of the dependent variable increases).
- The errors for different observations are independent and therefore uncorrelated (no autocorrelation of residuals).
- None of the X_i is perfectly correlated with any other X_i. If two or more variables were perfectly correlated, the equation could not be solved at all; since that would lead to a singular matrix that could not be inverted. Even if one or more pairwise correlations is very high although not unity, the parameter estimates are less likely to provide reasonable forecasts.

3.2 Uses and Misuses of $\overline{R}^2$

As already shown, the formula for $\overline{R}^2$ can be written as

$$\overline{R}^2 = 1 - \frac{\text{unexplained variance} * (n-1)}{\text{total variance} * (n-k)} \tag{3.3}$$

In the simple bivariate model, $k = 2$; so unless R^2 or n are quite low, there is very little difference between the two measures. However, when k increases as variables are added to the regression, the difference between R^2 and $\overline{R}^2$ can become quite large.

Table 3.1 Examples of R^2 and $\overline{R}^2$ under different circumstances.

R^2	n	k	$\overline{R}^2$
0.99	105	5	0.990
0.60	105	5	0.584
0.25	105	5	0.220
0.99	25	5	0.988
0.60	25	5	0.520
0.25	25	5	0.100
0.99	10	5	0.982
0.60	10	5	0.280
0.25	10	5	–0.350

3.2.1 Differences Between R^2 and $\overline{R}^2$

The practical significance of this adjustment means that $\overline{R}^2$ can never be increased by adding another variable if its t-ratio is less than unity. As $(n - k)$ becomes small – i.e., as the number of variables increases to the point where it is almost as great as the number of observations – the difference between R^2 and $\overline{R}^2$ increases dramatically; whereas when R^2 is very high and there are many degrees of freedom, the difference between the two measures is minimal. Also note that $\overline{R}^2$ can be negative, whereas that can never be the case for R^2. A few examples show how these two measures compare under different circumstances (table 3.1).

Note that if the t-ratio is greater than 1 for an additional variable, $\overline{R}^2$ will increase when that variable is added; if $t < 1$, it will decrease when it is added.

3.2.2 Pitfalls in Trying to Maximize $\overline{R}^2$

Most beginning – and even intermediate – model builders often follow the procedure of estimating several different regression equations and then choosing the one with the highest $\overline{R}^2$. What is wrong with that procedure?

To a certain degree, it would not be sensible to choose one equation over another just because it had a lower $\overline{R}^2$. However, there are several good reasons why maximizing $\overline{R}^2$ might not produce the best forecasting equation, some of which are as follows.

1 Adding lagged values of the dependent variable will invariably increase $\overline{R}^2$ for variables with strong trends, but it will almost certainly raise the error in multi-period forecasting.

2 Using what is essentially the same variable on both sides of the equation will also boost the fit, but will invariably result in worse forecasts.
3 Using seasonally unadjusted data will result in a better R^2, but only because you are "explaining" the seasonal factors that are already explained. In an extreme example, this is the "money supply causes Christmas" phenomenon.
4 In general, the stronger the trend, the higher the R^2, and the weaker the trend, the lower the R^2. But equations that merely explain a common trend provide no help in the forecast period.
5 Explaining an "extreme" value (one for which the error term is more than three standard deviations from the mean) by the use of dummy variables will always boost R^2, but will usually not provide any additional accuracy to the forecast.
6 Running hundreds of additional regressions to boost R^2 a little bit – "torturing the data until they confess" – may conceivably improve the forecast, but it is more likely that the slight improvement in the correlation is due to random elements that will not reoccur in the forecast period.

I have already noted that changing the form of the dependent variable may result in a lower $\bar{R}^2$ – but also a lower standard error, and hence a better forecast. That means maximizing $\bar{R}^2$ and minimizing the standard error of the estimate are not necessarily the same thing.

3.2.3 An Example: the Simple Consumption Function

To see this, consider various different estimates of the simple consumption function, where consumption is a function only of disposable income. The following example shows that shifting the dependent variable from consumption to saving reduces the R^2 but keeps the standard error (SE) unchanged. Also, shifting to a percentage first-difference equation sharply reduces the R^2 but also reduces the SE significantly. Clearly, maximizing R^2 is not the same thing as minimizing the standard error of the equation. All these equations are based on annual data.

A simple function in levels yields the result

$$\begin{aligned} CH = 43.1 + &1.11 * DIH - 0.27 * M2H \\ &(32.3) \qquad\quad (5.2) \\ \bar{R}^2 = 0.998;\ &SE = 53.9;\ DW = 0.63. \end{aligned} \tag{3.4}$$

The H indicates inflation-adjusted magnitudes, as distinguished from the same variables without the H, which are thus in current dollars. *M2* is the money supply. The numbers in parentheses under the *DIH* and *M2H* terms are the *t*-ratios. Since a *t*-ratio greater than 2 is generally considered significant, both variables appear to be extremely significant. Also, the R^2 appears to be extremely high.

To the first-time econometrician, this function, showing a near-perfect correlation between consumption and disposable income, might appear to be an excellent equation. In fact it is a very poor equation for predictive purposes, mainly because it is structurally deficient. An increase in income should raise consumption, but not by more than that increase in income; in this case the gain is 1.11 times as much. Second, the money supply may or may not be a very important determinant of consumer spending, but it certainly is not negative, as shown in this equation.

We can see the major changes that occur when the common trend is removed by using percentage changes. The R^2 drops sharply, but the coefficients make much more sense. The resulting equation is

$$\begin{gathered}\%\Delta CH = 0.86 + \underset{(8.8)}{0.58} * \%\Delta DIH + \underset{(5.8)}{0.24} * \%\Delta M2H \\ \bar{R}^2 = 0.777;\ SE = 0.84;\ DW = 2.10. \qquad (3.5)\end{gathered}$$

The $\bar{R}^2$ has decreased from 0.998 to 0.777, but the parameter estimates are much more reasonable. This equation says that every 1% increase in income will boost consumption by 0.58% the first year, while every 1% increase in the money supply will boost consumption by 0.24%. Again, both terms appear to be highly significant, and there is no autocorrelation of the residuals, so the goodness-of-fit statistics are not overstated.

The SE figures are not directly comparable because one is in levels and the other is in percentage changes. To compare these, we multiply the SE of 0.84 by the mean value of consumption over the sample period, which is 2579. For the levels equation, the standard error of estimate for consumption is 53.9, whereas for the percentage first-difference equation it is only 21.7 when converted to levels, or less than half as large. This finding can be verified by using EViews to print out the "forecast" values of consumption for both of these equations.

As an exercise, the student should estimate this equation using logs, first differences, first differences of logs, and the saving rate, which can be defined here as (income minus consumption)/income.[1] The summary statistics are given in table 3.2. In these examples, the constant terms have been omitted. The most important points to note are the following.

1 The goodness-of-fit statistics are about the same in logarithms as in levels, and the differences are very small. In this particular case, there has been no impact

[1] In practice, saving equals disposable income minus personal outlays, which equals consumption plus interest paid by persons plus personal transfer payments to the rest of the world. These minor differences need not concern us at this juncture.

Table 3.2 Summary statistics of the example.

Form of equation	Income coefficient	*t*-ratio	Money supply coefficient	*t*-ratio	R^2	*SE*	*SE* converted to levels	*DW*
Levels	1.11	32.3	–0.27	–5.2	0.998	53.9	53.9	0.63
Logs	1.01	23.9	–0.01	–0.2	0.998	0.0213	55.4	0.33
First differences	0.71	8.3	0.24	3.9	0.754	27.6	27.6	1.39
Percentage changes	0.58	8.8	0.24	5.8	0.777	0.87	21.7	2.10
Log first differences	0.58	8.8	0.24	5.8	0.777	0.0084	21.7	2.10
Saving levels	–0.11	–3.2	0.27	5.2	0.752	53.9	53.9	0.63

on the importance of the trend by switching to logarithms because both consumption and income have risen at about the same rate throughout the sample period. In other cases, however, logarithms can make a bigger difference, such as an equation for stock prices, which rose much faster than profits during the 1990s.

2 The first-difference and percentage change equations have much lower R^2, but also much lower *SE*. Also, the *DW* is much better, which means the goodness-of-fit statistics are not overstated. From a theoretical point of view, the income coefficient is also much more reasonable. In this case, removing the trend generates a much better forecasting equation, even though the fit appears to be much lower.

3 There is very little difference between the first differences and the percentage first differences. In the latter case, *SE* is a little lower and *DW* is a little better. In general, the percentage form is to be preferred when forecasting variables with strong trends because the size of the dependent variable is trendless instead of increasing over time.

4 The results for the percentage first differences and the first differences of logarithms are virtually identical. That will always be true because of the way logarithms are defined; it is not just a fluke for this set of equations.

5 Note that the R^2 is much lower for the saving levels equation than for the consumption levels equation. However, also note that the *SE* is identical. There has been a major reduction in R^2 even though the equation is the same (with the signs reversed), because saving as defined here is identically equal to income minus consumption. That alone should convince you not to rely exclusively on R^2 as a measure of how well your equation will forecast.

As we progress through the next few chapters, it will become clear that:

- the apparently high correlation shown in the levels equation often represents a common trend rather than a true behavioral relationship
- several variables are missing in the levels equation, as shown by the low *DW*
- significant autocorrelation and heteroscedasticity are present in the levels equation
- because consumption accounts for about two-thirds of GDP and hence income, to a certain extent the same variable occurs on both sides of the equation.

Real-life econometric equations are obviously much more complicated than this simple example. However, even at this level we can see that simply judging an equation by its R^2 and the coefficients by their t-ratios will often lead to very disappointing forecasts.

3.3 Measuring and Understanding Partial Correlation

Quite often it is the case that two variables X and Y will appear to have a very high correlation, but when one calculates a regression equation that includes a third variable Z, the partial correlation between X and Y will disappear. That means the correlation between the Y and the residuals based on a regression between X and Z is zero. For example, the number of marriages might be quite highly correlated with the number of drunken drivers arrested, but when population is added to the equation, that correlation disappears, since it only reflects a common trend.

While not as frequent, it can also happen that X and Y are uncorrelated, but when variable Z is added to the equation, both Y and Z became significant because they are negatively correlated with each other. For example, we might find a very low simple correlation between capital spending and interest rates, but when the growth in output is added to the regression, that term is significantly positive, while interest rates become significantly negative. That is because, in a statistical sense, high interest rates are usually associated with low growth, and vice versa.

3.3.1 Covariance and the Correlation Matrix

It is a simple matter to calculate the covariance matrix for all the variables used in a given regression, but that doesn't impart much useful information because the variables are generally of different magnitudes (e.g., some are interest rates, some are in billions of dollars, some are percentage changes, and so on). However, this defect can be easily remedied by transforming the covariance matrix into the correlation coefficient matrix using the formula given in chapter 2, which is repeated here:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}. \tag{3.6}$$

The correlation coefficient matrix can be easily observed in EViews by choosing a selected set of variables to be estimated in the equation, and then use the "Views" command to select this option.

3.3.2 Partial Correlation Coefficients

The correlation matrix shows which variables have strong positive and negative correlations with each other. However, more information is needed to make useful choices in a multiple regression equation. For that purpose we use the concept of *partial correlation*, which is the correlation between the dependent variable and a given independent variable *once the impact of all the other independent variables in the equation have been taken into account.*

The formulas become somewhat tedious without matrix notation for the general linear model, so we restrict them here to the case of an equation with two independent variables, X_2 and X_3.[2] If the dependent variable is Y, then

$$r_{YX_2X_3} = \frac{r_{YX_2} - r_{YX_3} \cdot r_{X_2X_3}}{(1 - r^2{}_{X_2X_3})^{1/2} \cdot (1 - r^2{}_{YX_3})^{1/2}}. \tag{3.7}$$

Let us see what this formula means. The denominator is included just for scaling purposes; i.e., it converts a covariance matrix to a correlation matrix, so we focus on the terms in the numerator. Suppose that r_{YX_2} were 0. It might appear, from looking at the simple correlation matrix, that X_2 would not belong in the equation. However, suppose that X_2 and X_3 have a strong negative correlation, and r_{YX_3} was significantly positive. In that case, the partial correlation between Y and X_2 *given that X_3 is also in the equation* would be significantly positive.

The thrust of these comments is that it is not sufficient to look at a variance–covariance or correlation matrix and simply choose those variables that have a high correlation with the dependent variables. In a multiple regression equation, all of the interactions must also be considered.

No one can be expected to see all these partial correlations right off the bat. Some trial and error in estimating regressions is always to be anticipated. Most model builders run an initial regression, look at the residuals, and then try to

---

[2] For further discussion, see Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th edn (Irwin McGraw-Hill, Boston), 1998, pp. 100–1; and Johnston, Jack, and John DiNardo, *Econometric Methods*, 4th edn (McGraw-Hill, New York), 1997, pp. 76–83. The latter reference also presents the k-variable case.

find some other variable which fits the unexplained residuals. There is nothing wrong with this procedure if it is not carried to extremes.

3.3.3 Pitfalls of Stepwise Regression

Given these comments, it might seem useful to have a regression program that chooses variables automatically. The model builder would select a fairly large set of all possible variables, the program would pick the one that initially had the highest correlation with the dependent variable and run that regression, then correlate the residuals with the variable that had the highest correlation with those residuals, run another regression, compare the residuals with the variable that had the highest correlation, and so on until no more significant variables were found. Such a program is known as *stepwise regression*, but in fact it doesn't work very well for several reasons. Indeed, the potential for misuse of partial correlation coefficients is probably greater than its potential usefulness.

The trouble with this approach, besides the obvious lack of any theory, is that it fails to take into consideration the possibility of negative covariance. It may be that neither variable X_2 nor X_3 has a significant correlation with Y, but when included in the same regression both of them are significant. As noted above, one key example might be capital spending as a function of output (or capacity utilization) and interest rates. The first-order correlations are not very strong, yet – as might be expected – there is a negative correlation between output and interest rates, especially when lags are taken into account. Mechanically using a stepwise regression package would miss this interaction. For reasons of this sort, the method is not recommended.

3.4 Testing and Adjusting for Autocorrelation

Since autocorrelation is primarily a problem that occurs in time-series (as opposed to cross-section) data, only time-series data are considered here, with the notation adjusted accordingly.

The general linear time-series model with autocorrelation can be written as

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_k X_{kt} + \varepsilon_t \text{ and } \varepsilon_t = \rho\varepsilon_{t-1} + v_t \tag{3.8}$$

where ρ is calculated as

$$\rho = \frac{\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\sigma_\varepsilon^{\;2}} = \frac{\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1})}{[\mathrm{Var}(\varepsilon_t)]^{1/2}[\mathrm{Var}(\varepsilon_{t-1})]^{1/2}}. \tag{3.9}$$

ρ is the correlation coefficient calculated for residuals in period t and $t - 1$.

3.4.1 Why Autocorrelation Occurs and What it Means

Autocorrelation is present in the vast majority of time series for the following reasons.

- *Errors of measurement*. Data reported by the government are based on incomplete information. Missing observations tend to be interpolated in such a way that the data are smoothed. Also, the same biases may occur systematically in the sampling program for different time periods.
- *Omitted variables*. This is probably the most serious problem. Positive correlation of residuals often signifies one or more significant variables are missing. In some cases, such as expectational variables, these missing variables cannot be measured precisely.
- *Misspecification of existing variables*. This often refers to a nonlinear equation. Sometimes the equation is piecewise linear, which means the underlying structural relationship has shifted during the sample period. The equation is linear in both periods but one or more of the coefficients has changed. In other cases, the coefficients vary with the phase of the business cycle. Sometimes one or one or more of the independent variables should be raised to some power other than unity. That could mean an exponential power such as squared, cubed, etc., or it could indicate an inverse correlation, where the form of the independent variable should be $1/X$ instead of X.
- *The effect of habit*. Even if the data are correct and the equation is correctly specified, economic decisions are often based on habit, so that the error term in this period after taking into account all the relevant variables really is correlated with the error term in the previous period. Sometimes this information can be used to help improve forecast accuracy, but using it runs the risk of generating poor forecasts whenever habits do change.

The presence of autocorrelation does not affect unbiasedness or consistency, but it does affect efficiency. That means the standard errors that are calculated using OLS are understated, and hence the significance levels of the individual terms (and often the entire equation) are overstated. Results that appear to be significant actually are not. I emphasize that the presence of autocorrelation in the residuals does not necessarily mean the parameter estimates are incorrect; it simply means that the sample period statistics will probably understate errors during the forecast period.

In working with time-series data, it is often the case that *quarterly* data will indicate significant autocorrelation, while the same equation estimated with *annual* data shows no autocorrelation. If this happens, and the coefficients in the annual equation are approximately the same, it is generally better *not* to include the autoregressive adjustment in the quarterly equation when using it for forecasting purposes. Such a result often means that the quarterly data are artificially smoothed, not that variables are missing. If on the other hand some of the coefficients in the corresponding annual equation become insignificant, the quarterly equation needs more work before it can be used for forecasting.

3.4.2 Durbin–Watson Statistic to Measure Autocorrelation

The Durbin–Watson (DW) statistic is used in statistics almost as widely as R-bar squared.[3] It is defined as

$$DW = \sum (\varepsilon_t - \varepsilon_{t-1})^2 \Big/ \sum \varepsilon_t^2 . \tag{3.10}$$

This formula has a marked resemblance to the formula for ρ. We can expand the numerator and write

$$DW = \left(\sum \varepsilon_t^2 + \sum \varepsilon_{t-1}^2 - 2\sum \varepsilon_t \varepsilon_{t-1}\right) \Big/ \sum \varepsilon_t^2 . \tag{3.11}$$

In large samples, $\Sigma\varepsilon_t^2$ is approximately equal to $\Sigma\varepsilon_{t-1}^2$, so

$$DW = 2\left(\sum \varepsilon_t^2 - \sum \varepsilon_t \varepsilon_{t-1}\right) \Big/ \sum \varepsilon_t^2 = 2(1-\rho). \tag{3.12}$$

Thus if $\rho = 0$, DW is 2; if $\rho = 1$, $DW = 0$; and if $\rho = -1$, $DW = 4$.

The significance level of DW for 95% confidence is generally around $DW = 1.4$, although it is lower for smaller samples and higher for larger samples. There is also an upper and lower bound; within this range, the results are indeterminate.

[3] For further discussion, see Pindyck and Rubinfeld, pp. 160–9; and Johnston and DiNardo, pp. 179–84. The original articles are Durbin, J. and G. S. Watson, "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37 (1950), 409–28; and 38 (1951), 159–78. The tables for significant upper and lower bounds were extended in Savin, N. E., and K. J. White, "The Durbin–Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica*, 45 (1977), 1989–96. Modifications for the test when the equation contains no constant term are given in Farebrother, R. W., "The Durbin–Watson Test for Serial Correlation when there is no Intercept in the Regression," *Econometrica*, 48 (1980), 1553–63. Adjustments for this statistic when the lagged dependent variable is used in the equation are found in Durbin, J., "Testing for Serial Correlation in Least Squares Regression when some of the Regressors are Lagged Dependent Variables," *Econometrica*, 38 (1970), 410–21.

Since the significance level depends on the sample size and the number of independent variables, as well as containing this indeterminate range, another way to check the significance level is to run the regression with the AR(1) adjustment and see whether that coefficient is significant. If *DW* is above 1.6, that test is not necessary.

3.4.3 Autocorrelation Adjustments: Cochrane–Orcutt and Hildreth–Lu

It is straightforward to calculate ρ by calculating the underlying regression and determining its value from the residuals, using the formulas given above. Such a value can then be inserted into the underlying equation, with the equation rewritten as

$$(Y - \rho Y_{-1}) = \beta_1(1-\rho) + \beta_2(X_1 - \rho X_{1-1}) + \beta_3(X_2 - \rho X_{2-1}) + \ldots + (\varepsilon_t - \rho\varepsilon_{t-1}) \qquad (3.13)$$

where the value of ρ is as calculated in (3.6). If we let $v_t = \varepsilon_t - \rho\varepsilon_{t-1}$, then v_t can be tested for autocorrelation, and a second iterative value of ρ can be obtained. This method can be repeated for as many iterations as desired until convergence is reached; generally it takes fewer than ten iterations.

This method is known as the *Cochrane–Orcutt* method,[4] and is by far the most popular method of adjusting for autocorrelation. In EViews this is listed as the AR(1) transformation, and is implemented by reestimating an equation with autocorrelation by adding the term AR(1) to the equation. It is not necessary to respecify each term individually.

Under unusual circumstances, though, it is possible that the initial ρ selected might not be the only one; there could be multiple values for ρ. For that reason, the Hildreth–Lu method[5] is sometimes used. This scans the values of ρ from –1 to +1 in a grid with units of, say, 0.1. This method then finds the global maximum and then zeros in on this value using a grid of, say, 0.01, until the answer is found to the desired degree of precision. However, this method isn't used very much because Cochrane–Orcutt usually gives the same answer and is easier to implement.

Sometimes ρ is very close to unity, in which case the equation reduces to a first-difference equation. *A useful rule of thumb is to use first differences or percentage changes if the value of the DW statistic is less than R^2.*

For example, if $\bar{R}^2$ were 0.99 for a levels equation (not unusual, as we have seen), and *DW* for that equation were 0.50 (also not unusual), the equation

---

[4] Cochrane, D., and G. H. Orcutt, "Applications of Least-Squares Regressions to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, (1949), 32–61.

[5] Hildreth, G., and J. Y. Lu, "Demand Relations with Autocorrelated Disturbances," *Michigan State University Agricultural Experiment Station Technical Bulletin* 276, November, 1960.

should probably be reestimated in percentage first differences. On the other hand, if *DW* were greater than 1.0, it would probably be better to stay with the levels equation and try to improve the equation by experimenting with lag structures or adding more variables. Other methods of trend removal are considered later in the book.

3.4.4 Higher-Order Autocorrelation

So far we have considered only the possibility of first-order autocorrelation. The error term might also be correlated with own values lagged more than one time period, so that in the regression

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_k\varepsilon_{t-k} + v_t \tag{3.14}$$

the various ρ_i, $i > 1$, would also be significant.

It is unlikely, although not impossible, that a whole string of the ρ_i would be significant, but that is not usually what happens. Instead, we find that in time-series data there is some autocorrelation with residuals a year ago; so the term ρ_4 would be significant for quarterly data – but not the intervening ρ's. For monthly data, the term ρ_{12} would be significant – but not the intervening ρ's.

If this sort of situation does arise, there are four possible treatments. One is seasonally adjusting the data (which can be easily done in EViews) and see whether the problem disappears. If it does not, the three remaining options are as follows:

- Add the lagged variable with the 4(quarter) or 12(month) lag.
- Use the SAR term in EViews, which means calculating the fourth (or 12th order) autocorrelation factor but none of those in between, unless they are explicitly specified.
- Switch to a percentage first-difference equation, using the current period minus four quarters (or 12 months) ago.

If seasonal autocorrelation appears in the data, it is usually preferable to seasonally adjust the data or, if the data are robust enough, use percentage first-difference equations. Firms are generally interested in how well their orders or sales are doing compared with a year ago, rather than comparing them with periods that have different seasonal factors.

3.4.5 Overstatement of *t*-Ratios when Autocorrelation is Present

Under reasonable values of ρ, the standard errors can be understated by more than 80%, thus providing grossly inaccurate measures for goodness-of-fit statistics. For the simple bivariate case $Y_t = \beta X_t + \varepsilon_t$, the variance of β in an equation with autocorrelated residuals can be as large as $(1 + r\rho)/(1 - r\rho)$, where ρ is the autocorrelation coefficient and r is the correlation coefficient of x_t and x_{t-1}. The proof is found in Maddala.[6]

In series with strong time trends, it would not be unusual to find that ρ is 0.8 or higher, and r is 0.9 or higher. Using these values, we would find that the ratio given above is 1.72/0.28, or 6.14. That means the stated t-ratios could be more than six times as high as the "actual" estimates. Hence a variable with a t-ratio of 10 under the above autocorrelation conditions might not even be significant at the standard 5% level. However, this is an upper limit. In most cases, positive autocorrelation of the residuals does not bias the standard errors down by the full amount. In relatively large samples (50 to 150 observations), the bias is usually between one-half and two-thirds of the amount given by this formula. Nonetheless, this formula should serve as a warning about the overstatement of t-ratios when significant positive autocorrelation is present.

In addition to overstating the t-ratios, it is likely that the existence of autocorrelation means the equation is misspecified, in which case trying to reduce autocorrelation by the use of the AR(1) adjustment will not improve forecasting accuracy. There is actually a fairly simple test for this. Consider the two equations

$$Y_t = \rho Y_{t-1} + \beta X_t - \beta\rho X_{t-1} + \varepsilon_t \tag{3.15}$$

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 X_t + \alpha_3 X_{t-1} + \varepsilon_t. \tag{3.16}$$

These are the same with the restriction that $\alpha_1\alpha_2 + \alpha_3 = 0$. EViews contains a simple test to determine whether that relationship holds or not, known as the *Wald coefficient restriction*. If the relationship does not hold, then the AR(1) adjustment should not be used, and further experimentation with the form of the equation is warranted.

Autocorrelation does not bias the coefficients unless the lagged dependent variable is used. However, it does overstate their significance, which means the forecast error will probably be larger than indicated by the sample period statistics for the equation.

3.4.6 Pitfalls of Using the Lagged Dependent Variable

A quick glance at the least-squares printouts will reveal that R^2 is much higher, and the standard error of estimate is much lower, when the lagged dependent variable is included, either explicitly or with an AR(1) adjustment.

---

[6] Maddala, G. S., *Introduction to Econometrics*, 2nd edn (Prentice Hall, Englewood Cliffs, NJ), 1992, pp. 241–4.

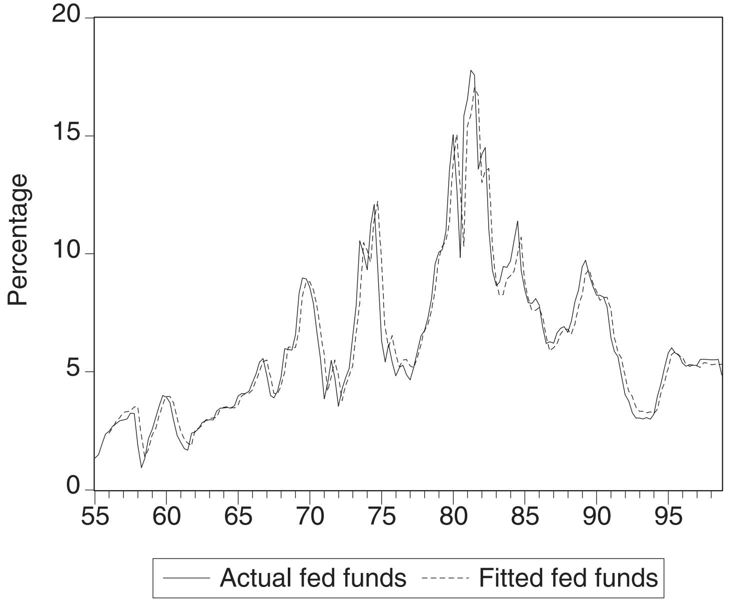

Figure 3.1 Fitted Federal funds rate as a function of inflation, and actual funds rate lagged one quarter.

For those who want to forecast only one period in advance, the use of all relevant information, including the lagged dependent variable, may improve forecast accuracy. This possibility is discussed in greater detail in Part III. However, a higher R^2 and *DW* statistic closer to 2 does not mean the multi-period forecasting accuracy of the equation has been improved by adding the lagged dependent variable. Indeed, beyond one period in the future, forecasting accuracy is often *diminished* by adding the lagged dependent variable as an additional independent variable.

To see the pitfalls involved, consider an equation in which the Federal funds rate is a function of the rate of inflation over the past year. This equation does not work very well because Fed policy was far different in the 1970s, under Arthur Burns and G. William Miller, than it was in the 1980s and 1990s, under Paul Volcker and Alan Greenspan. An equation in which the Fed funds rate is only a function of the inflation rate thus gives unsatisfactory results: it explains only about half the variance, the DW is an unsatisfactorily low 0.25, and the standard error is over two percentage points.

Now suppose the regression is recalculated by adding the funds rate lagged one quarter. On the surface, the results look much better: $\overline{R}^2$ has risen from 0.58 to 0.91, *DW* has improved from 0.21 to 1.51 (indicating no significant autocorrelation of the residuals), and the standard error has been cut in half.

However, upon closer examination, the equation misses all the turning points by one quarter (see figure 3.1). Maybe it seems as though the fitted values "catch up" after one quarter, but that is only because they depend on the actual lagged value. If the fitted lagged value were to be used, the simulated values would drift ever-further away from the actual value.

This is perhaps an extreme example, but it illustrates the point well. Using the lagged dependent variable on the right-hand side of the equation will often result in an equation with apparently superb sample period fits, but it will be useless for forecasting because the lagged dependent variable also has to be predicted. Even for single-period forecasting, the equation given above would miss virtually every turning point, and hence would be useless for actual forecasts.

Forecasting errors often arise when trying to predict Y if the most important independent variable is Y_{-1}, which means relying primarily on that variable for predicting Y. By definition, Y_{-1} never turns down ahead of time. Thus relying on the lagged dependent variable means missing almost all the turning points. Furthermore, Y_{-1} will not only rise during the first period that Y fell, but will continue to rise because the forecast will continue to contain erroneous feedback from the variable that failed to turn around. The more periods that are predicted with this equation, the worse the forecasts.

Thus it seldom if ever pays to put the lagged dependent variable on the right-hand side of an equation that will be used for multi-period forecasting. If that variable continues to be very significant in spite of all other changes that have been made, switch to percentage first differences, or other methods that eliminate the trend, many of which are discussed in detail in the next chapter.

3.5 Testing and Adjusting for Heteroscedasticity

The other major reason for non-normality of residuals is that, the larger the value of the dependent variable, the larger the value of the residual. In essence that means giving more weight to the larger sample points in the regression equation, which means that the goodness-of-fit statistics are overstated, just as is the case for autocorrelation.

Heteroscedasticity can arise from two major sources. The first is that as the size of the dependent variable increases, the absolute size of the error term increases, even though the percentage error does not rise. That happens primarily in cross-section data. The second cause stems from outliers, which can be handled either with dummy variables or by omitting those variables from the regression. Cross-section data are considered first.

3.5.1 Causes of Heteroscedasticity in Cross-Section and Time-Series Data

In cross-section data the causes and cures of heteroscedasticity are fairly straightforward, and can best be illustrated by an example. Suppose someone is estimating data from a panel survey of consumers, whose income ranges all the way from (say) \$10,000 to \$1,000,000. Let us assume for these purposes that the same factors govern consumption at all levels of income, so the various consumption functions are similar. In that case, the standard deviation for the \$1 million consumers would be 100 times as large as the standard deviation for the \$10,000 consumers. A few rich consumers would therefore dominate the sample in terms of the statistical tests.

The straightforward solution to this problem is to scale the results so that (say) a 5% error for the rich gets the same weight as a 5% error for the poor. That can be done using ratios or weighted least squares. However, as was pointed out earlier, most of the emphasis in forecasting is on time-series analysis. How does heteroscedasticity arise in those cases?

The simplest case stems from the fact that most time series increase over time (consumption, production, employment, prices, etc.). If the level of the dependent variable is, say, ten times as great at the end of the sample period as it was at the beginning, then the error term is also likely to be ten times as great. However, if this variable is correlated with an independent variable (income in the consumption function) with the same general trend, the residuals probably will not be heteroscedastic. Even if heteroscedasticity remains, this might have the net result of giving more recent observations greater weight, which in many cases is a good idea anyhow.

The other problem, which is not related to the trend, is one of extreme values. This can perhaps best be illustrated by looking at financial data. On Monday, October 19, 1987, the Dow Jones Industrial Average plunged 508 points, or 22% – a decline almost twice as much as the next largest percentage drop (including Black Tuesday in 1929). If all percentage changes are treated equally in a statistical sense, the results will be biased in the sense that the market will be shown to fall more on the 19th of each month, or each Monday, or each October.

In macroeconomic data, price equations might be dominated by energy shocks, thereby neglecting the importance of other key variables such as unit labor costs, capacity utilization, monetary policy, and so on.

These problems can be quite severe in the sense that not only are the goodness-of-fit statistics overstated, but the parameter estimates themselves will become biased, and forecasts based on these estimates will invariably generate the wrong answers. Case Study 3 illustrates how this can happen.

3.5.2 Measuring and Testing for Heteroscedasticity

Researchers generally want to neutralize the distorting influence of extreme outliers without excluding them from the sample period entirely, assuming that they contain some relevant information. Using dummy variables essentially takes them out of the equation, although it overstates the goodness-of-fit statistics; omitting them entirely from the sample period will not distort these statistics. The usual procedure is to test the residuals to see whether heteroscedasticity is present.[7]

Unlike autocorrelation, there is no standard test such as the DW statistic. The two most common tests are the ARCH LM test (which stands for Autoregressive Conditional Heteroscedasticity – Lagrangian Multiplier) suggested by R. F. Engle,[8] and the White test, developed by Halbert White.[9]

The ARCH LM test is based on the comment made above that the most recent residuals are likely to be larger in typical time-series analysis. Engle thus suggests estimating the equation

$$\varepsilon_t^2 = \beta_1 + \beta_2\varepsilon_{t-1}^2 + \beta_3\varepsilon_{t-2}^2 + \ldots + \beta_k\varepsilon_{t-k-1}^2 \tag{3.17}$$

where the researcher picks the value of k (EViews asks you to supply this number). If the equation is significant, as measured by the F-ratio, then heteroscedasticity exists. The White test is a more general one and involves expanding the regression equation to include the value of the variables squared, and the cross-products among independent variables. Here too the F-ratio is used to determine whether the equation is significant.

Under the assumptions of the classical linear model, heteroscedasticity will affect the standard errors and goodness-of-fit statistics but not the parameter estimates. In other words, OLS (ordinary least squares) estimates are still unbiased even in the presence of heteroscedasticity. The most common adjustment to the variance/covariance matrix, and hence to the standard errors of an estimate, is known as White's correction. Another method, known as the Newey–West correction, gives consistent estimates even if both autocorrelation and heteroscedasticity are present (which is likely to be the case in standard time-series analysis). Generally the results do not vary much between the two methods.

While these two tests are theoretically sound, they still are not entirely satisfactory in the sense that, if a distortion or bias is introduced into the parameter estimate by an extreme outlier, these corrections will not fix the problem. These adjustments, like virtually all statistical tests, have been developed under the assumption of normal distributions, whereas in fact the existence of an extreme outlier indicates that the distribution is *not* normal.

While the t-ratios are a bit more realistic after applying these adjustments, one does not really get to the root of the problem of heteroscedasticity by apply-

[7] For further discussion, see Pindyck and Rubinfeld, pp. 146–57; and Johnston and DiNardo, pp. 162–7.

[8] Engle, Robert F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50 (1982), 987–1008.

[9] White, H., "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48 (1980), 817–38.

ing White's correction, or the Newey–West correction. There is nothing wrong in adjusting the standard errors and it should be done if heteroscedasticity is found to be present, but most of the time it does not make much difference. By experimenting, the reader can quickly determine that using these adjustments generally has much less effect on the parameter estimates than using the AR(1) adjustment, or changing from levels to changes.

Another method of diminishing heteroscedasticity is to use weighted least squares or put the dependent variable in ratio form. Weighted least squares (WLS)[10] is somewhat arbitrary in the sense the model builder must choose which weights to use; one common method is to use one of the independent variables that has the same trend or scale factor as the dependent variable. In the consumption function, for example, that would be income, whether time-series or cross-section data were being used. One could also take the ratio of consumption/income as the dependent variable; the impact of these changes is discussed in the following chapter. In most cases, using WLS does not change the parameter estimates very much either.

If one of the values of a time series has an extreme outlying value, thus leading to heteroscedasticity, it cannot be ignored; otherwise that one value will dominate the results, and the equation will eventually be reduced to fitting these one or two extreme points. One way to proceed is usually to treat this with a dummy variable, as discussed in chapter 4. Another is simply to disregard the errant observations completely. This process, often known as masking, consists of automatically excluding any observation where the error term is larger than a preassigned multiple of the standard error of the overall equation. One common rule of thumb is to exclude observations whose error term is more than three times the standard error.

In many cases, omitting outliers will generate parameter estimates that will produce forecasts with smaller errors than if the outliers were included. However, simply omitting all observations that cannot be explained can become a dangerous procedure. In particular, one should determine whether those outliers were caused by a specific exogenous development. For example, electric power usage would rise more during extremely cold winters or hot summers, insurance claims would rise dramatically after a hurricane, and entertainment expenditures for a given city – especially one of moderate size – would rise sharply after the local baseball team wins the World Series. At first glance it might appear these outliers should be discarded, but in fact they can usefully be correlated with the indicated exogenous development. Only in cases where outlying values do not appear to be related to any realistic independent variable should they be excluded from the equation.

---

[10] For further discussion see Pindyck and Rubinfeld, pp. 148–9; and Johnston and DiNardo, pp. 171–2.

3.6 Getting Started: An Example in EViews

To provide a preview of many of the issues that occur in building forecasting models, a simple equation for annual constant-dollar department store sales is now presented. This series starts in 1967; thus for the percentage change equation, the first data point is 1968. The level of department store sales is a function of real disposable income less transfer payments, stock prices as measured by the S&P 500 index, and the availability of credit is measured by two variables: the amount of consumer credit outstanding, and the yield spread between long- and short-term rates, which measures the willingness of banks to make loans to consumers. The estimated values are

$$\begin{aligned}
&\text{SALES} \\
&\quad = \underset{(2.3)}{2.736 + 2.49} * YDHXTR + \underset{(5.1)}{8.78} * SP + \underset{(2.2)}{0.142} * YLDSPRD(-1) + \underset{(2.4)}{7.24} * CRED \\
&\qquad RSQ = 0.992 \; DW = 0.42.
\end{aligned} \tag{3.18}$$

This is the standard form in which equations will be presented in this text. The numbers in parentheses under the coefficients are t-ratios. *RSQ* is R^2 adjusted for degrees of freedom. *SALES* is the level of constant-dollar department store sales. *YDHXTR* is real disposable income excluding transfer payments. *SP* is the S&P 500 stock price index. *YLDSPRD*(–1) is the difference between the Aaa corporate bond yield and the Federal funds rate lagged one year. *CRED* is the amount of consumer credit outstanding.

While this might appear to be an excellent equation, with an adjusted R^2 in excess of 0.99, most of the correlation is due to the upward trend; an equation in which the log of sales is a function only of a time trend would explain 98% of the variance. Also the residuals shown in figure 3.2 indicate this equation does a poor job of tracking changes in sales. The low *DW* of 0.42 indicates some structural defect. For this reason, as is so often the case in time-series variables with strong trends, we consider the percentage change equation

$$\begin{aligned}
&\%SALES \\
&\quad = 1.19 + \underset{(3.6)}{0.592} * YDHXTR + \underset{(2.6)}{0.068} * \%SP + \underset{(3.2)}{0.673} * YLDSPRD(-1) \\
&\qquad RSQ = 0.685; \; DW = 1.58.
\end{aligned} \tag{3.19}$$

This equation has a much lower adjusted R^2, but is a better structural equation. The graph of the residuals, as shown in figure 3.3, shows far less autocorrelation. Also, the consumer credit term has dropped out of the equation;

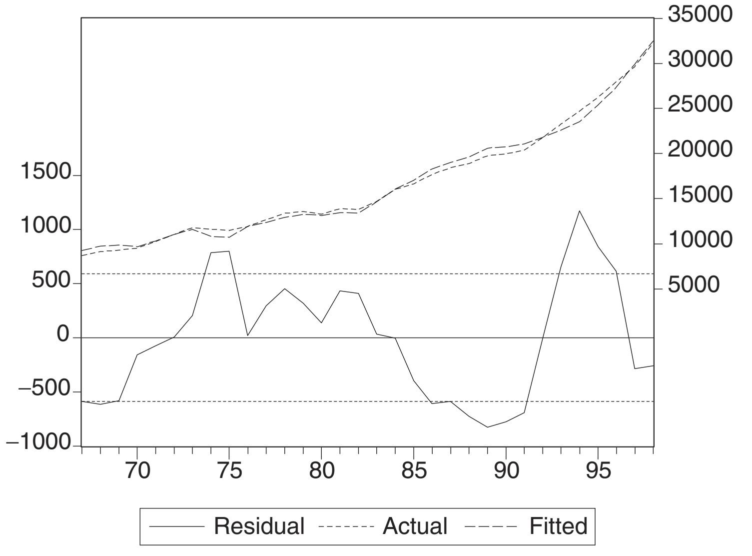

Figure 3.2 Example – see the text.

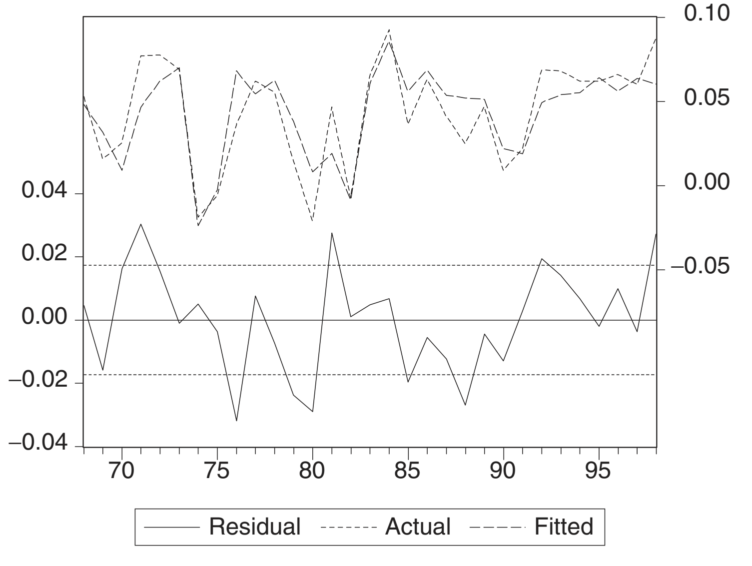

Figure 3.3 Example – see the text.

indicating it may have reflected reverse causality – consumer credit outstanding rose because department store sales rose, not the other way around. This is a fairly typical example of how an equation with a much lower R^2 is more likely to provide accurate forecasts.

Case Study 1: Predicting Retail Sales for Hardware Stores

One of the most important aspects of practical business forecasting is to provide actual examples and show how the equations were developed. In many cases, the problems that occur are as important as the successes, so the point of these case studies is to illustrate pitfalls as well as show impressive results. In general, each chapter from this point on will include three case studies. This chapter presents case studies where autocorrelation or heteroscedasticity play an important role in determining the equation that is actually chosen for forecasting.

The dependent variable in the first case study is retail sales at hardware and building materials stores in constant dollars. In general, hardware sales, like other types of consumer purchases, are related to real disposable income (*YDH*); this category of sales is also sensitive to the level of housing starts (*HST*). In addition, a reduction in interest rates, measured here by the corporate bond rate (*FAAA*), will boost home maintenance and additions, and a decline in the unemployment rate (*UN*) will also boost construction. In this example quarterly data are used; as is the case for department store data, they start in 1967.1. The equation with these four variables and their appropriate lags in quarters is

$$
\begin{aligned}
&HDWSALES \\
&\quad = -2.375 + \underset{(84.8)}{2.78^*\, YDH} - \underset{(7.2)}{0.153^*\, FAAA(-3)} + \underset{(8.7)}{0.770^*\, HST} - \underset{(3.1)}{0.093^*\, UN(-1)} \\
&\qquad RSQ = 0.983;\ DW = 0.25.
\end{aligned}
$$

All the variables appear to be significant, and *RSQ* seems quite high, but the *DW* statistic is far too low at 0.25. There are two standard ways to handle this: use the AR(1) adjustment or use a percentage first-difference equation.

With the AR(1) adjustment, the reader can verify that all the coefficients remain significant, *RSQ* reportedly rises to 0.996, and *DW* is slightly above 2. Nonetheless, an examination of the residuals shows that such an equation invariably misses the turning points.

Most hardware store owners (or manufacturers of hardware equipment) are interested in sales relative to year-earlier levels, which suggests taking the percentage change from year-earlier levels and treating the independent variables in the same fashion. When this is done, all four variables remain significant, but housing becomes relatively more important and income relatively less important, since the spurious correlation from the common trend is no longer present. In the equation below, the 4 at the end of each variable means a four-quarter change; % is percentage change; and Δ is first difference (used for variables without trends). Actual instead of percentage first differences are often used for variables without trends.

$$\begin{aligned}
&\%HDWSALES, 4 \\
&\quad = 1.70 + 0.997 * \%YDH, 4 - 1.56 * \Delta FAAA(-3), 4 \\
&\qquad\qquad (6.3) \qquad\qquad\qquad (6.0) \\
&\quad\; + 0.132 * \Delta HST, 4 - 1.84 * \Delta UN(-1), 4 \\
&\qquad (14.2) \qquad\qquad\qquad (5.9) \\
&\qquad\qquad RSQ = 0.846;\; DW = 0.91.
\end{aligned}$$

All variables remain highly significant, and *DW* is slightly better at 0.91, although it still indicates significant positive autocorrelation. The histogram shows no significant heteroscedasticity. When the AR(1) term is used, all the coefficients remain significant and *DW* rises to 1.85.

Which equation is likely to give the better forecasts? The standard error of the percentage first-difference equation is 2.8% of the mean value, which is slightly more than $7 billion, which indicates a standard error of about $200 million. By comparison, the standard error of the levels equation is $325 million. On this basis the percentage change equation has a substantially smaller standard error.

Comparing this equation with one containing the AR(1) term (not shown here) shows very little difference in the value of the coefficients. That is not surprising in the sense that, according to standard statistical theory, the presence of autocorrelation biases the *t*-ratios but does not usually distort the estimates of the coefficients themselves, although as shown below that is not always the case.

The reader may also wish to calculate the same equation using annual data and verify that (i) the coefficients do not change very much, and (ii) there is no significant autocorrelation when annual data are used.

Since the coefficients are almost the same, either equation will generate almost the same forecasts. In this case, then, it does not matter which form of the equation is used. However, that is not always true, as seen in the next case study.

Case Study 2: German Short-term Interest Rates

We have already seen that the real Federal funds rate varies quite significantly depending on who is chairman of the Federal Reserve Board. It might be interesting to see how much real German short-term interest rates vary (the three-month bill rate is used, symbol TB3GER), depending on who is heading the Bundesbank.

To see this, regress the German short-term rate on the German rate of inflation and the growth in the German money supply (M2); the latter term reflects Bundesbank (BBK) policy that when money supply growth accelerates, the BBK tightens. The US Federal funds rate lagged one quarter is also included because US interest rates tend to dominate world financial markets. The US inflation rate is also included with a negative sign to show that the *real* Federal funds rate is more important: when US interest rates rise because of general worldwide inflation, the BBK is less likely to tighten than when the Fed raises real rates.

$$\begin{aligned} &TB3GER \\ &\quad = -0.799 + \underset{(9.2)}{0.711} * GERINFL + \underset{(4.7)}{0.106} * \%GERM2,4 + \underset{(12.0)}{0.531} * FFED(-1) \\ &\qquad - \underset{(4.4)}{0.238} * INFL(-1) + \underset{(3.1)}{0.086} * \%GERM2(-4),4 + \underset{(6.3)}{0.602} * GERINFL * DUNIF \\ &\qquad\qquad RSQ = 0.847;\ DW = 0.79. \end{aligned}$$

The prefix GER before inflation and M2 money supply indicate those variables are for Germany. *DUNIF* is a dummy variable for unification explained below. The term *%GERM2,4* means percentage change in the German money supply over the past four quarters.

This may seem like a reasonable equation, but once again *DW* is low. An examination of the residuals shows that German interest rates were well above their predicted values in the early 1990s. That occurred for a very specific reason: after unification, inflation rose because of questionable government policies that artificially equalized the value of the Dmark and the Ostmark, which should have been set no higher than half the value of the DM. The net effect was that unemployment rose dramatically in the former East Germany, leading to a massive increase in transfer payments and higher inflation. Since the inflation was due to political bungling, the BBK tightened more than previously when inflation rose.

That suggests using a dummy variable that is 0 before unification and 1 afterwards (*DUNIF*), and multiplying that variable by the rate of inflation. When that is tried, such a variable is highly significant and the fit improves substantially in later years. However, *DW* still indicates positive autocorrelation of the residuals, so the AR(1) adjustment is tried again.

This time, however, the value of the coefficients changes markedly. The dummy variable completely drops out of the equation, and so does the US inflation rate. Perhaps the latter variable is not appropriate after all, but the impact of inflation after unification would seem to be a theoretically justified variable. However, the result indicates otherwise once the AR(1) adjustment is included.

The histogram of the residuals shows that heteroscedasticity is present both with and without the AR(1) term. The White or Newey–West algorithms to adjust the variance–covariance matrix hardly changes the results; in this case, the presence or absence of heteroscedasticity is not very important. This leaves the main question: should the dummy variable term be used in the forecasting equation?

$$TB3GER$$

$$= 1.59 + 0.632{*}\,GERINFL + 0.065{*}\,\%GERM2,4 + 0.303{*}\,FFED(-1) - 0.090{*}\,INFL(-1)$$

$$\quad (4.0) \qquad\qquad (1.9) \qquad\qquad (4.3) \qquad\qquad (0.9)$$

$$+ 0.037{*}\,\%GERM2(-4),4 - 0.075{*}\,GERINFL{*}\,DUNIF + 0.865{*}\,AR(1)$$

$$\quad (1.2) \qquad\qquad (0.2) \qquad\qquad (15.6)$$

$$RSQ = 0.917;\ DW = 1.74.$$

No one rule will always apply to this situation, but we can draw some specific conclusions in this case. First, in the equation with the dummy variable, the coefficients indicate that, after unification, a 1% rise in the inflation rate would boost short-term interest rates by 1.3%, whereas in the equation with AR(1) the coefficient is only 0.6%. On an a-priori basis, a coefficient of about 1% would seem reasonable. Second, the reaction of the BBK to the inept political decisions following unification is a one-time event that is unlikely to be repeated. It thus appears that neither version of the equation is totally satisfactory.

It is often the case that specific dummy variables designed to fit a few data points will disappear once the AR(1) adjustment is incorporated. In that sense, using the AR(1) term will often help the forecaster reduce the amount of "curve fitting" that is inherent in any econometric estimation procedure. Here, the AR(1) adjustment warns us off using a dummy variable that is highly significant but would probably reduce forecast accuracy. However, that does not necessarily mean the equation with AR(1) will forecast better; it probably means the equation needs to be respecified. More examples of this appear later in the text.

Case Study 3: Lumber Prices

The histogram for the percentage change in lumber prices is given in figure 3.4. Note that the changes are far from being normally distributed; severe heteroscedasticity exists because of a few outlying observations.

Classical statistical assumptions state only that the residuals of the equation should be normally distributed; that criterion does not have to apply to the original series itself. Nonetheless, the point of this example is to show what happens when we try to fit an equation that has severe outlying values. Sometimes the result is nonsensical.

To start, estimate an equation in which the percentage change in lumber prices is a function of housing starts, the change in the capacity utilization rate for manufacturing, lagged changes in the value of the dollar, changes in the price of oil, and a dummy variable for the imposition and termination of wage and price controls. That is not a very good equation in terms of the goodness-of-fit statistics, and explains only a little more than one-third of the total variance.

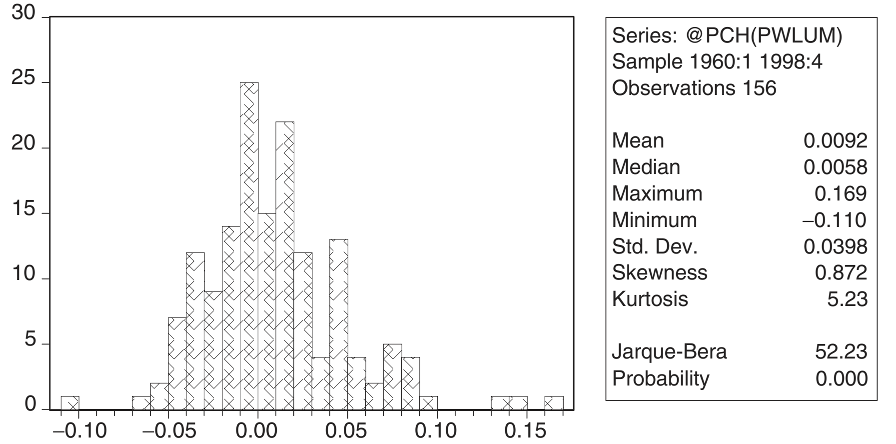

Figure 3.4 Case study 3 – see the text.

$$\begin{aligned}
&\%PPI\ LUMBER \\
&\quad = -6.31 + \underset{(5.5)}{0.474}{*}\,HST(-2) + \underset{(4.3)}{0.000478}{*}\,\Delta CUMFG\hat{}2 \\
&\quad\quad - \underset{(2.8)}{0.083}{*}\,\Delta DOLLAR(-8),4 + \underset{(3.0)}{0.058}{*}\,\%POIL(-1) + \underset{(4.5)}{5.55}{*}\,DUMWP \\
&\qquad\qquad RSQ = 0.356;\ DW = 1.50.
\end{aligned}$$

The residuals in figure 3.5 show that the price of lumber rose far more than is explained by the equation in 1980.1, and fell far more than is explained in 1980.2. Once again, there is a very specific reason. Paul Volcker became Fed chairman in late 1979; his predecessor, G. William Miller, appeared to be accommodating higher inflation, which led to runaway speculation in commodities. When Volcker tightened and imposed credit controls in early 1980, the speculative binge immediately collapsed. Since events of this sort had never occurred before, they are not drawn from the same sample. The Jarque–Bera test shows the residuals are not normally distributed.

Given this explanation, what should be done about the forecasting equation: add a dummy variable for this one period, eliminate those two observations, or add an economic variable that attempts to explain this abrupt shift in sentiment?

Adding a dummy variable improves *RSQ* but does not help the predictive accuracy of the model. Neither does eliminating these two points completely. Suppose the enterprising researcher scours the data for another series that would seem to measure the change in sentiment – which turns out to be the change in the price of gold. When that variable is added to the equation, the fit improves materially, and the residuals are now normally distributed, as shown

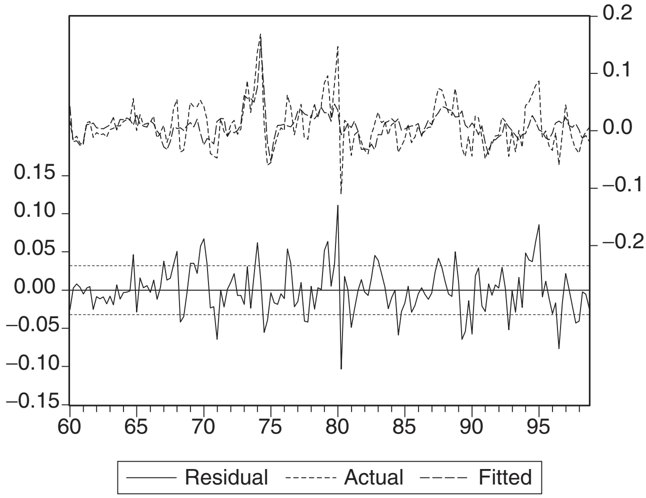

Figure 3.5 Case study 3 – see the text.

in figure 3.6. One could perhaps argue that the price of gold accurately reflects speculative fever.

$$\begin{aligned}
&\%PPI\ LUMBER \\
&\quad = -4.92 + \underset{(4.7)}{0.378^{*}\,HST(-2)} + \underset{(4.3)}{0.000511^{*}\,\Delta CUMFG\text{^}2} - \underset{(3.4)}{0.091^{*}\,\Delta DOLLAR(-8),4} \\
&\qquad + \underset{(2.2)}{0.038^{*}\,\%POIL(-1)} + \underset{(4.8)}{5.40^{*}\,DUMWP} + \underset{(5.6)}{0.049^{*}\,\Delta PGOLD} \\
&\qquad\qquad RSQ = 0.464;\ DW = 1.39.
\end{aligned}$$

However, the first half of 1980 was the only time that the change in the price of lumber was correlated with the change in the price of gold, which was really a proxy variable for runaway inflationary expectations. That can be tested by reestimating this equation starting in 1983.1, after inflation had returned to low levels; both that term and the change in oil prices drop out.

The price of gold actually has no economic relationship with the price of lumber. What happened, however, is that by estimating an equation with a few extreme outlying values – i.e., with severe heteroscedasticity – the least-squares formula gives a disproportionately high weight to these few values. Thus any other variable that has similar peaks and troughs, whether really related or not, will appear to be highly significant because it happens to fit those few periods. In recent years, the price of lumber and the price of gold have moved in opposite directions, and for good reason: when inflation is low, interest rates decline and housing starts rise, hence boosting the price of lumber. Yet low inflation also leads to a decline in gold prices. Hence using a positive correlation between

Chapter 4

Additional Topics for Single-equation Regression Models

Introduction

The previous chapter discussed the issues of autocorrelation and heteroscedasticity in the residuals, and illustrated the standard statistical adjustments that are used when these problems arise. It also pointed out the possible pitfalls of building forecasting models when those conditions are present.

Multicollinearity occurs when two or more of the independent variables in the regression equation are very highly correlated. Unlike autocorrelation and heteroscedasticity, there are no specific tests for multicollinearity, but it can be even more serious because it distorts the values of the coefficients themselves, rather than affecting only the goodness-of-fit statistics. The usual problem is that while the sum of the coefficients of highly collinear variables is close to the true underlying value, the individual coefficients contain significant errors. Thus unless the relationship between these values is exactly the same in both the sample and forecast periods, predictions from such a model are likely to contain serious mistakes.

The problems of multicollinearity stem from two major sources: *different* variables that are highly collinear, and *lagged values of the same variable*, which will be highly collinear if that series contains strong trends. The treatment of these causes is quite different. In the first case, variables should be combined, or the strong common trend should be removed by using first differences, percentage changes, ratios, or weighted least squares. In the second case, several lagged values of the same variable should be combined into only a few terms by the use of distributed lags. Both these methods are considered in detail.

4.1 Problems Caused by Multicollinearity

In equations where several variables have with strong time trends – consumption, investment, prices, sales, production, income, etc. – they are likely to be highly collinear. However, there is no explicit test that determines when this problem will distort the results. The equations must be examined on an individual basis. I will indicate how extreme multicollinearity can distort the parameter estimates, and the best way to reduce if not entirely eliminate this problem.

To see how multicollinearity can give ridiculous results, consider the admittedly far-fetched example in which consumption is regressed against:

- disposable income
- the major components of disposable income: wages, transfer payments, taxes, and all other components of personal income
- wages, transfer payments, and total income. (There is some double counting here, which is on purpose.)
- wages, transfer payments, taxes, other income, and total income. (This matrix ought to be singular, except we have introduced some rounding errors so it will convert. But since total income equals the sum of the first three variables, the results can reasonably be expected to be nonsensical.)
- wages, transfer payments, taxes, and other income, but all in percentage changes, so multicollinearity is no longer a problem.

The summary statistics for these regressions are as follows:

$$\begin{aligned} C = -21.4 + &\underset{(348.5)}{0.923 * YD} \\ R^2 = 0.999;\ SE &= 32.8;\ DW = 0.23. \end{aligned} \tag{4.1}$$

$$\begin{aligned} C = -20.5 + \underset{(7.8)}{0.556 * W} + \underset{(9.8)}{1.212 * TR} + \underset{(9.4)}{1.223 * YOTH} - \underset{(1.8)}{0.364 * TX} \\ R^2 = 0.999;\ SE = 25.6;\ DW = 0.47. \end{aligned} \tag{4.2}$$

$$\begin{aligned} C = -10.2 - \underset{(1.4)}{0.210 * W} - \underset{(9.8)}{0.440 * TR} + \underset{(8.1)}{0.975 * YD} \\ R^2 = 0.999;\ SE = 28.4;\ DW = 0.38. \end{aligned} \tag{4.3}$$

$$\begin{aligned} C = -18.3 - \underset{(1.0)}{40.95 * W} - \underset{(1.0)}{41.57 * TR} - \underset{(1.0)}{41.60 * YOTH} - \underset{(1.0)}{40.36 * YD} - \underset{(1.0)}{40.93 * TX} \\ R^2 = 0.999;\ SE = 25.6;\ DW = 0.67. \end{aligned} \tag{4.4}$$

$$\begin{aligned} \%C = 0.013 + \underset{(5.3)}{0.591 * \%W} + \underset{(3.5)}{0.087 * \%TR} - \underset{(2.1)}{0.066 * \%TX} + \underset{(3.3)}{0.206 * \%YOTH} \\ R^2 = 0.784;\ SE = 0.0113;\ DW = 1.91. \end{aligned} \tag{4.5}$$

All terms are in current dollars; all R^2 are adjusted for degrees of freedom. The numbers in parentheses are t-ratios. C is consumption, YD is disposable income, W is wages, TR is transfer payments, $YOTH$ is other personal income, and TX is personal income taxes.

Note that the overall goodness-of-fit statistics in the first four equations are all virtually the same when the components of income are disaggregated. That ought to be a tipoff something is wrong, for theory suggests that at least the short-term marginal propensity to consume from volatile components of income is smaller than from stable components. The fact that the estimated value of the coefficients for *TR* and *YOTH* are greater than unity in equation (4.2) looks suspicious right away. In equation (4.3), the negative signs on *W* and *TR* are clearly inappropriate. In equation (4.4), with almost complete multicollinearity, the results are nonsensical – even though the R^2 stays the same.

When multicollinearity is eliminated by taking percentage changes, wages are the most important variable, and transfers are much less important because the major cyclical component of transfers is tied to the business cycle, and rises when other income declines. Later we will see that this consumption function is still seriously incomplete because no monetary variables are included. But you can't tell that from the statistics, since *DW* is 1.91.

Also note that the standard errors become enormous when extreme multicollinearity is present. That doesn't always happen; but when it does, that is an obvious hint this condition exists. In that case, the logical choice is to drop one or more of the variables.

There are a few tests that suggest multicollinearity is present, but they are not discussed here – nor are they included in EViews – because (i) they do not provide any additional information that cannot be gleaned from the correlation matrix and comparison of the sizes of the standard errors, and (ii) unlike with autocorrelation and heteroscedasticity, there is no simple way to fix the problem. There are some tests known as "complaint indicators," which tell you that multicollinearity is present, but not what to do about it.

If an equation with extreme multicollinearity is used for forecasting, the results will contain very large errors if there is even a tiny change in the relationship between the multicollinear independent variables, because the coefficients have been blown up to unrealistically high values. Ordinarily, if the relationship between the independent variables changes a little bit, the forecast error will be quite small. Hence it is generally a poor idea to generate forecasts using an equation with extreme multicollinearity.

The following lessons can be drawn from the above example:

1 Extreme multicollinearity will often result in nonsensical parameter estimates, including the wrong signs, and unusually high standard errors.
2 Rearranging the *same data* in different linear combinations will often reduce the degree of multicollinearity.
3 If the levels form of the equation is retained, the degree of multicollinearity can be reduced by dropping one or more of the variables.
4 The best way to solve the problem is to remove the common trend, either through percentage first differences or other methods, which are discussed next.

Table 4.1 Empirical results for the example.

Form of equation	Current income	Lagged income	Current price	Lagged price	$\bar{R}^2$	*SE*	*DW*
Level	0.200	0.025	7.0	12.2	0.985	21.7	0.52
	(2.9)[a]	(0.4)	(0.9)	(1.7)			
Logarithm	1.77	0.41	–0.1	0.4	0.993	0.15	0.37
	(2.5)	(0.6)	(–0.2)	(0.6)			
Percentage change	1.40	1.12	–0.63	–0.36	0.675	0.041	1.73
	(4.3)	(3.5)	(–4.1)	(–2.4)			
Deviations from trend	1.85	0.86	–0.64	–0.36	0.941	0.039	0.60
	(4.8)	(2.4)	(–3.4)	(–1.9)			
Ratio	0.083[b]	0.083	–0.0061	–0.0011	0.956	0.0060	0.27
	(1.4)	(1.4)	(–3.0)	(–0.6)			
Weighted least squares	0.207	0.032	9.8	11.8	0.995	21.2	0.67
	(3.4)	(0.5)	(1.2)	(1.6)			

[a] The numbers in parentheses are *t*-ratios.
[b] Constant term.

4.2 Eliminating or Reducing Spurious Trends

There are several common methods for removing trends from the data in regression equations with time-series data. Each of these methods is discussed, followed by a list of the major advantages and disadvantages of each method. The results will be illustrated by empirical estimates of an equation for airline travel for each of these five cases. While the results from this example are fairly typical, varying results may be obtained for different functions. The point of the airline travel function is to illustrate how these various methods reduce multicollinearity and forecast error, not to define a set of rules that will work in all cases.

Case Study 4: Demand for Airline Travel

In this case study, an equation is estimated to explain the demand for airline travel using several different methods to reduce or eliminate the trend. In these equations, the dependent variable is revenue passenger miles traveled; the data are taken from the website of the Air Transport Association of America, www.air-transport.org. Income is real disposable personal income, and price is the price of airline fares measured in cents per mile divided by the CPI. This equation is first estimated in linear form, and then reestimated for the following five cases (the empirical results are shown in table 4.1):

- log-linear transformations
- percentage first differences
- deviations around trends
- ratios
- weighted least squares.

One would of course expect that airline travel is positively correlated with income, with an elasticity of greater than unity, and negatively related to the relative price of airline travel. However, the percentage change equation is the only one in which both current and lagged airline travel prices are significantly negative. In particular, in the levels equation, both price terms are positive, which makes no sense. The logarithm equation is only slightly better in this regard. However, note that when the trends are removed from the logarithm equation, it is much improved.

Also note that the results for weighted least squares are almost the same as those for the unweighted levels equation in spite of the strong upward trend in airline travel. The weights in this case are the values of disposable income, but approximately the same results would have been obtained with other similar weighting factors. That is a common finding; using this option seldom makes much difference.

The coefficients in the percentage change equation are elasticities, so they can be examined in terms of economic relevancy. The combined income terms show an elasticity of 2.52, suggesting that airline travel is a highly discretionary good that increases sharply during years of prosperity. The price elasticity is –1.00, an interesting finding for the following reason. The marginal costs of adding an additional passenger to the flight are close to zero; therefore economic theory would say that maximizing profit occurs at about the same price as maximizing revenue, which happens at a price elasticity of unity. The fact that the total result turns out to be exactly –1.00 is a coincidence, but it is nonetheless revealing to find this regression indicates airlines have priced their product at a point that maximizes profits.

4.2.1 Log-linear Transformation

- *Main advantage:* removes some of the common trend, dampens outlying values
- *Secondary advantage:* coefficients easily interpreted as elasticities
- *Major disadvantage:* autocorrelation generally remains just as serious a problem
- *Secondary disadvantage:* implies underlying function has constant elasticities, which may not be the case
- *Related methods:* levels

Consider the form of the airline travel function

$$\log TR = \beta_1 + \beta_2 \log Y + \beta_3 \log Y_{-1} + \beta_4 \log P + \beta_5 \log P_{-1} \tag{4.6}$$

where TR is airline travel and P is the relative price of airline travel, which has fallen an average of 3% per year over the sample period. There is nothing particularly wrong with this equation, except in levels form it would be

$$TR = e^{\beta 1} Y^{\beta 2} Y_{-1}^{\beta 3} P^{\beta 4} P_{-1}{}^{\beta 5}. \tag{4.7}$$

In fact, the underlying equation might or might not be multiplicative. There are no a-priori rules for determining when an equation is linear and when it is log-linear. In a log-linear equation, the elasticities remain the same over the entire sample period. That may or may not be an appropriate assumption.

A linear demand curve means that, at relatively low prices, the demand is inelastic, so an increase in price will boost total revenues; while at relatively high prices the demand is elastic, so a further increase in price will reduce total revenues. The log-linear demand curve assumes that the price elasticity is constant along the entire length of the curve. On an a-priori basis there is no way to determine which assumption is better. In the airline equation, both the levels and the logarithm equations give non-significant results for the price terms, so the issue cannot be decided with these equations.

Cost (or supply) curves are usually flat in the region of constant returns to scale, and then start rising at an increasing rate as diminishing returns set in. In that case, the function is neither linear nor log-linear, and must be estimated using nonlinear techniques, a method discussed later in this chapter. Production functions are generally thought to be log-linear, with constant elasticities of substitution. It is often assumed a certain percentage increase in costs results in the same percentage increase in prices whether the economy is in a boom or a recession. As a matter of fact, the change in the markup factor is probably due more to monetary policy and expectations than to the phase of the business cycle, leading to a complicated nonlinear relationship that usually is not estimated directly (i.e., a linear approximation is used by including monetary factors separately). But here again there is no conclusive empirical evidence that using logarithms is better or worse.

In many cases, the empirical evidence does not permit one to choose between linear and log-linear equations. If the theory provides strong reasons to expect constant elasticities, use the logarithmic formulation; otherwise use the linear form. For series with strong trends, the results generally do not differ very much. The logarithm form is often preferred because, as noted above, the coefficients are elasticities, making comparison easier if one is working with equations involving hundreds of commodities, countries, companies, or individuals.

As noted in chapter 2, using percentage changes is virtually the same as using first differences of logarithms, although there is one slight difference: in calculating percentage changes, there is some ambiguity about whether the denominator should be the current period, the previous period, or some average of the two periods. By taking differences of logarithms, this ambiguity is resolved.

However, except in unusual cases, the difference between the two choices of equations is minuscule.

4.2.2 Percentage First Differences

- *Main advantage:* eliminates all traces of trend
- *Secondary advantage:* eliminates "imbalances" between levels and rates
- *Major disadvantage:* one or two outlying points, which might not make much difference in a levels equation, could distort the regression estimates
- *Secondary disadvantage:* may obscure long-run relationships
- *Related methods:* first differences (without the percentage) gives similar results, although over a long time period many first differences also contain a significant trend; also, since the percentage change coefficients are elasticities, they are easier to interpret; first differences of logarithms give essentially identical results

The equation for annual percentage changes for airline travel yields robust parameter estimates, as noted above. One point of interest is the much lower $\overline{R}^2$, which is just one more example of how that statistic can often be misleading. The standard error of the equation is 4.1%, and the mean value of the dependent variable is 270, which means in levels terms the error is 11.1. That is much lower than the standard error of 21.7 for the levels equation.

While this equation works well, percentage changes in quarterly or monthly data are often distorted by seasonal quirks and random factors. Hence the noise overwhelms the signal and the results are not robust. This problem is sometimes handled by taking percentage differences over the same quarter the previous year, which finesses the seasonal problem. The trouble with this method cannot be detected in a single-equation approach, but creates problems when these equations are combined in a simultaneous model with multi-period forecasts. Using annual percentage first differences sometimes creates a spurious two-year cycle in the forecasts, especially when reinforced by other equations of the same form. This problem can be reduced if not totally eliminated by using a four-quarter moving average, lagged one quarter, instead of the four-quarter lag itself. Thus we have

$$\Delta Y/Y_{-1} = \beta \Delta X/X_{-1} \text{ (often too much noise)} \tag{4.8}$$

$$(Y - Y_{-4})/Y_{-4} = \beta (X - X_{-4})/X_{-4} \text{ (could cause spurious cycles)} \tag{4.9}$$

$$(Y - \overline{Y})/\overline{Y} = \beta (X - \overline{X})/\overline{X} \tag{4.10}$$

where $\overline{X} = (X_{-1} + X_{-2} + X_{-3} + X_{-4})/4$ (preferred if equations are to be used in simultaneous-equation models).

4.2.3 Ratios

- *Main advantage:* where applicable, eliminates trends
- *Secondary advantage:* series are "smooth" and not dominated by outliers; also, long-run parameters are better developed, and equation is "balanced" if untrended series are included
- *Major disadvantage:* must be sure the ratio of the dependent variable does not have a trend, or will not have a trend in the forecast period, otherwise the problem of spurious trends could remain.
- *Secondary disadvantage:* constant term may introduce nonlinearities
- *Related methods*: weighted least squares (see section (4.2.5))

We start with the levels function

$$TR = \beta_1 + \beta_2 Y + \beta_3 Y_{-1} + \beta_4 P + \beta_5 P_{-1} \tag{4.11}$$

where TR is airline travel, Y is disposable income, and P is relative price of airline travel.

In levels form this equation could have several drawbacks. In particular, Y and Y_{-1} could be collinear, and the effect of P might be swamped by the trends. One could take percentage first differences, but sometimes that understates the long-run change in the dependent variable due to a change in income or prices. This equation can be reformulated as

$$TR/Y = \beta_1 + \beta_2\, Y_{-1}/Y + \beta_3 P + \beta_4 P_{-1}. \tag{4.12}$$

The relative price terms have not been divided by Y because that would introduce a spurious trend at least as serious as the one we are trying to eliminate.

Until now the constant term has not been discussed, since it is not very important. However, in this case, the β_1 term has some economic significance, since the actual forecasts will be calculated by multiplying the entire equation by Y, so β_1 becomes the coefficient of the income term. Thus in its linear form, this equation has no constant term at all.

To compare this equation to the levels and percentage change forms, it is useful to calculate the elasticities. The sum of the coefficients of the two income terms, which happen to be identical, is 0.166. Since the mean value of TR/Y is 0.071, and the mean value of Y_{-1}/Y is approximately unity, then the income elasticity is 0.166/0.071, or approximately 2.34 – quite similar to the 2.52 value contained in the percentage change equations.

The mean value of airline prices (cents/mile in constant dollars) is 12.8, so the price elasticity (combining terms) is –0.0072 * 12.8/0.071, or about –1.3. That is slightly higher than the –1.0 figure obtained from the percentage change equations. The standard error in levels form is obtained by multiplying the *SE* of the equation, which is 0.006, times the mean value of real disposable income, which is 3,355, yielding a comparable figure of 20.1. That is slightly lower than the levels equation *SE* of 21.7 but substantially higher than the comparable percentage change equation *SE* of 12; so on balance, the percentage change equation would appear to be a better choice.

4.2.4 Deviations Around Trends

- *Main advantage:* by definition, trends are eliminated; don't have to worry about trends remaining, as in ratios
- *Secondary advantage:* a backup if percentage changes or ratios do not work
- *Major disadvantage:* in many cases, does not eliminate or reduce autocorrelation
- *Secondary disadvantage:* changes in trend in the dependent but not the independent variables could result in poorly performing equation
- *Related methods:* none

Once again consider the airline travel function in equation (4.11), but this time remove the trend from each variable. Representing the trend value for TR as TR_{TR} and using similar notation for the other variables, we can write

$$(TR - TR_{\mathrm{TR}}) = \gamma_1 + \gamma_2(Y - Y_{\mathrm{TR}}) + \gamma_3(Y_{-1} - Y_{\mathrm{TR}}) + \gamma_4(P - P_{\mathrm{TR}}) + \gamma_5(P_{-1} - P_{\mathrm{TR}}) \tag{4.13}$$

where in each case the trend variable is calculated by running a separate *regression equation* of the form $\log X = \alpha_1 + \alpha_2 t$ (where t is a time trend) for TR, Y, and P. In other words, these are logarithmic trends. This equation is equivalent to

$$\begin{aligned}\log TR \\ &= \gamma_1 + \gamma_2 \log Y + \gamma_3 \log Y_{-1} + \gamma_4 \log P + \gamma_5 \log P_{-1} + (1 - \lambda_2 - \lambda_3 - \lambda_4 - \lambda_5)T_{\mathrm{TR}}\end{aligned} \tag{4.14}$$

where T_{TR} is the time trend for airline travel, and the λ are some linear combination of the γ and the trend rates of trend growth in Y and P. If the sum of the λ's is not approximately equal to zero, then the method of deviations around the trend can reintroduce a spurious trend back into the equation, hence offsetting some of the detrending that this method is designed to accomplish.

For this reason, the use of differences around trends will not always work. Yet in many occasions it is useful, especially when percentage changes in the data are erratic or look like a random series, or when the relevant ratios still contain strong trends.

To estimate this equation, first estimate a regression of the log of travel on a time trend, and then calculate the residuals. This process is repeated for income and for the relative price of airline travel. In EViews, then estimate *RESID1* (the difference between the log of travel and its trend value) as a function of current and lagged *RESID2* (the difference between the log of income and its trend value), and current and lagged *RESID3* (the difference between log of relative price and its trend)

The results are quite instructive. Unlike with the logarithm equation, where the combined price elasticity was only –0.13 and clearly not significantly different from zero, we find that the price elasticity in this equation (measured by the coefficients, since this is a logarithmic equation) is exactly −1.00, the same result obtained for the percentage change equation. The income elasticity is 2.71, slightly higher than the 2.52 figure in the percentage change equation.

The next issue is how to compare standard errors; this equation, which is in logarithms, has a standard error of 0.039, compared with 21.7 in the levels equation. The standard error must be converted from logarithms to levels to draw any meaningful comparison which is done as follows. The mean value of the airline travel variable is 270.5; the natural log of that number is 5.600. One standard deviation from that number is 5.639; the antilog of that is 281.2. Hence the standard error of this equation converted to levels if 10.7, only about half the *SE* of the levels equation. Hence in this regard, the deviations-from-trend equation is superior.

The other factor to note is that *DW* is only 0.60, which suggests the possible use of the AR(1) transformation. When that happens, the equation improves in the sense that the income elasticity falls to 2.43; the price elasticity remains at −1.00. While *DW* now shows no autocorrelation, that is not a valid test with AR(1). The major change is that the income elasticity is now lower.

In this case, virtually the same result is obtained as occurs in the percentage change equation; that does not always happen. Also, note that while the logarithm equation does not give reasonable results, the equation is substantially improved when the trends are removed. Hence this equation is yet another case where strong trends in time-series data obscure the underlying result.

4.2.5 Weighted Least Squares

- *Main advantage:* best method for treating heteroscedasticity, although its use is more common for cross-section than time-series data
- *Major disadvantage:* essentially reduces equation to ratio form, which may still leave strong trends in key variables
- *Secondary disadvantage:* in most cases, arbitrary choices of weights
- *Related methods:* ratios

There are two cases: one where the variances are known, and the other where they are estimated. In the first case, the variables in the regression are simply divided by their respective variances. In practical terms, though, this information is hardly ever known.

In the more usual case, the changes in the variances are approximated by the changes in one of the trend variables (in the airline travel function, the obvious choice would be income). The equation is then divided through by that variable. In this case, weighted least squares has some similarity to ratios, but the results are usually closer to the OLS equations than the ratio equation.

There are few differences between the OLS and WLS equations because all of the independent variables have significant trends: income rises and relative price falls. When some of the independent variables do not have any trends, such as percentage changes or interest rates, WLS often improves the coefficients of these trendless terms, in which case the forecasting accuracy of the equation generally improves. Most of the time, however, there is not much difference between OLS and WLS estimates.

4.2.6 Summary and Comparison of Methods

Any time one is calculating regressions using time series with strong trends – whether they are components of aggregate demand and income, individual demand and supply functions, production functions, money supply, stock prices, or any other variable that grows over time – the original set of equations, based on the relevant theory, will usually show positive autocorrelation of the residuals, and will usually suffer from multicollinearity as well. These maladies could be due to a number of different factors, but most of the time the culprit is the strong common trend.

Both the sample period statistical results and the forecasting properties of the equation are likely to be unsatisfactory unless these problems are resolved. Using the lagged dependent variable on the right-hand side of the equation – or using an autocorrelation adjustment – will provide a "quick fix" in the sense that the sample period statistical tests will appear to be better, but often the multi-period forecasting record will become worse. It is better to use one of the methods mentioned here to eliminate the common trends.

Building econometric forecasting equations can never be reduced to a "cookbook" technique, and different problems call for different solutions, but the general checklist should be helpful in deciding which form of the equation to use.

1. It is usually easy to tell if strong trends are dominant simply by looking at the data. If not, check the correlation matrix. If a wide variety of different forms of the regression equation routinely yield $R^2 > 0.99$, it is reasonably clear all you are doing is measuring a common trend, and it is best to reduce or remove it.
2. Another hint will be a very low *DW* statistic; if *DW* is lower than R^2, that is usually a reliable signal that one should use percentage changes (or first differences of logarithms, which will give essentially the same result).
3. In equations with strong trends, high R^2, and low *DW*, percentage changes should be tried. The major drawback occurs if the resulting series is dominated

by random fluctuations, which usually means the absolute values of the parameter estimates are biased down. If monthly or quarterly data are being used, one logical choice is to try annual percentage changes (i.e., this month or quarter over the same month or quarter a year ago). For forecasting with multi-equation models, it is better for the lagged variable to be an average over the past year than simply a year ago, in order to avoid spurious cycles in forecasting more than one year out.

4 If the annual percentage change method does not work, consider either ratios or detrending each series. These methods often do not solve the problem of autocorrelation. Yet while an AR(1) adjustment will superficially solve that problem, it generally will not improve forecasting accuracy for multi-period predictions and often makes the errors larger.

5 For variables without trends – interest rates, inflation rates, foreign exchange rates of the dollar, etc. – levels equations are preferred. There may still be some autocorrelation, but that is best handled by improving the specification of the equation rather than by moving to percentage changes.

4.3 Distributed Lags

So far we have looked at the problem of multicollinearity as it applies to two or more independent variables with strong trends. For illustrative purposes we used annual data. However, an even more common cause of multicollinearity occurs when quarterly or monthly data are used and the theory suggests several lagged values of one or more of the independent variables. For example, consumption depends on lagged as well as current income. Because of multicollinearity, the estimated coefficients in regression equations will generally be nonsensical if an entire string of lagged variables is entered in a single equation. Yet theory does not tell us precisely how long the lag will be, nor what shape the distribution will take: whether 90% of the reaction will take place in the first time period, or whether it will be spread over several years.

In most key macroeconomic equations – consumption, investment, exports, interest rates, wages and prices, etc. – economic choices depend on lagged as well as current variables. The problem is obviously more important the shorter the time period considered: lagged values are more important for quarterly and monthly data than for annual data. At the industry level, changes in shipments, new orders, and inventories depend on what has happened in the past as well as the present. Only in cross-section data are lags generally considered unimportant.

4.3.1 General Discussion of Distributed Lags

Most of the time, the researcher must make some a-priori assumption about the shape and length of the lag distribution, otherwise the empirical testing can quickly get out of hand. As a general rule, the following four types of lag distributions are the most common in estimating regression equations.

The simplest kind of lag is the *arithmetic moving average with equal weights*. All terms have the same weight up to a certain point, beyond which all the weights are zero. For example, changes in wage rates this quarter might depend on changes in the inflation rate over the past four quarters with equal weights, followed by zero weights for longer lags. One could make a plausible case for such an assumption in the case where wages are changed only once a year, while price changes are continuous. In that case we might write

$$\%(wr) = \beta_0 + \beta_1[\%(cpi) + \%(cpi)_{-1} + \%(cpi)_{-2} + \%(cpi)_{-3}]/4 \tag{4.15}$$

where "%" is the percentage first difference operator; i.e., $\%(x) = (x - x_{-1})/x_{-1} * 100$.

This may not sound like a very realistic lag distribution, but could occur if, say, wage bargains incorporate all the inflation that has occurred since the previous contract was signed, but none of the inflation before that point. This is sometimes known as the "one-horse shay" assumption, based on the concept that a machine (originally a horse) would perform faithfully and at the same efficiency during its lifetime, and then, when it suddenly expired, wouldn't work at all. Most of the time, though, the weights on lag distribution do not suddenly become zero, so lag distributions are used that describe the actual situation more accurately.

The second type of lag distribution is a *declining weight moving average*, which can be either arithmetic or geometric. An example of an arithmetic distribution is:

$$(4Y_{-1} + 3Y_{-2} + 2Y_{-3} + Y_{-4})/10. \tag{4.16}$$

An example of a geometric distribution is

$$(0.8Y_{-1} + 0.64Y_{-2} + 0.512Y_{-3} + 0.4096Y_{-4})/\Sigma \tag{4.17}$$

where Σ is the summation of the weights used in the lag distribution.

The drawback to estimating the equation

$$C_t = \beta_1 + \beta_2\left(\sum_{i=1}^{k} \lambda^i Y_{-i}\right) \tag{4.18}$$

is that it requires a nonlinear estimation technique involving simultaneous estimation of both β_2 and λ (k is set by the researcher). As shown later, such results often fail to give satisfactory results; in addition, the significance levels of the parameter estimates can be tested only on a linear approximation. This equation can also be transformed into

$$C_t = \beta_2 Y + \lambda C_{t-1} \tag{4.19}$$

but estimating that equation means using the lagged dependent variable, which generally is not recommended.

The third type of lag distribution is an *inverted U* (or V). As the name implies, the weights start near zero, rise to a peak, and then tail off again. A common example would be the lag between orders and deliveries: immediately after the order is placed only a few goods are delivered, then the proportion rises, and eventually falls off to zero again. To a certain extent this lag structure is similar to the normal distribution. A typical example of an inverted U distribution is

$$\begin{aligned} Y_t = \beta_0 + \beta_1(X_t + 2X_{t-1} + 3X_{t-2} + 4X_{t-3} + 5X_{t-4} \\ + 4X_{t-5} + 3X_{t-6} + 2X_{t-7} + X_{t-8})/25. \end{aligned} \tag{4.20}$$

The fourth type of lag distribution is a *cubic distribution*. The initial weights are relatively large, then decline sharply, but then rise again before eventually declining to zero. This would occur in the case where an initial impact caused some variable to change; after that impact had worked its way through the economy, there would be a secondary impact, smaller but still significant. This lag distribution is often found in functions for capital spending: the initial plans are made based on variables with long lags, but then modified as more recent economic conditions change, especially when orders are canceled or construction is halted.

By now we are reaching the point of diminishing returns, since the researcher is basically specifying the overall lag structure without testing beforehand to see whether the assumption is a reasonable one. Hence a more general approach is needed. While one could certainly invent other, more complicated lag distributions, the exercise soon reduces to curve fitting and data mining rather than econometrics.

All the distributions described here have weights that eventually go to zero. There is no point in considering a lag distribution so long that the weights never go to zero. It doesn't really make any economic sense, and it couldn't be estimated anyhow. However, it is generally not known on an a-priori basis whether the weights are large or small at the beginning of the lag period, or whether they decline monotonically. A more generalized lag structure that can be estimated in the linear regression model is needed to fill this gap.

4.3.2 Polynomial Distributed Lags

This section discusses polynomial distributed lags; the geometric lag is a special case. The lag structure can indeed be specified as a geometrically declining lag. However, it could also turn out to be the shape of an inverted U or V, it could have two or more peaks, or the weights could first decline and then rise again. In the case of the lag between orders and contracts, and deliveries or completions of capital goods, the weights of the distribution would probably be relatively small at first, peak in the middle, and eventually decline to zero again. The concept of polynomial distributed lags, which was introduced into economics to estimate the lag structure between ordering and delivery of capital goods by Shirley Almon,[1] permits the user to choose the following values:

- whether or not the lag is constrained to zero at the near end
- whether or not the lag is constrained to zero at the far end
- the degree of the polynomial: linear, quadratic, cubic, etc.
- the total length of lag.

The PDL method is quite general: the user need only choose the parameters given above. On the other hand, the results are not always easy to interpret, since they often give very similar goodness-of-fit statistics for wide variations in length of lag and degree and shape of the polynomial. I will first explain the method, then give some standard examples. Since the general case involves a fair amount of tedious algebra, the example of a third-degree polynomial with a five-period lag and no endpoint restrictions is given next.[2] The lag specification is

$$Y_t = \beta_0 + \beta_1(w_0X_t + w_1X_{t-1} + w_2X_{t-2} + w_3X_{t-3} + w_4X_{t-4}) + \varepsilon_t \tag{4.21}$$

where

$$w_i = \gamma_0 + \gamma_1 i + \gamma_2 i^2 + \gamma_3 i^3,\ i = 0,1,2,3,4. \tag{4.22}$$

If the equation were a higher order than cubic (say, quartic), then equation (4.22) would have an additional term for i^4.

The simplest case is a linear polynomial – i.e., a straight line. If the distribution were constrained at the far end, and the lag were (say) five quarters, the underlying lag distribution would be a declining straight line that intersected the x-axis after five quarters. For a geometric distribution, the specification would call for a quadratic polynomial constrained at the far end.

After a fair amount of arithmetic, which involves substituting the w_i of (4.22) into the original equation (4.21), it can be shown that

[1] Almon, Shirley, "The Distributed Lag Between Capital Appropriations and Expenditures," *Econometrica*, 33 (1965), 178–96.

[2] For further discussion, see Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th edn (Irwin McGraw-Hill, Boston), 1998, pp. 236–8. Also see Maddala, G. S., *Introduction to Econometrics*, 2nd edn (Prentice Hall, Englewood Cliffs, NJ), 1992, pp. 423–9).

$$\begin{aligned}Y_t = \beta_0 &+ \beta_1 c_0 (X_t + X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}) \\ &+ \beta_1 c_1 (X_{t-1} + 2X_{t-2} + 3X_{t-3} + 4X_{t-4}) \\ &+ \beta_1 c_2 (X_{t-1} + 4X_{t-2} + 9X_{t-3} + 16X_{t-4}) \\ &+ \beta_1 c_3 (X_{t-1} + 8X_{t-2} + 27X_{t-3} + 64X_{t-4}).\end{aligned} \quad (4.23)$$

Since the product of two parameters cannot be estimated with linear methods, it is generally assumed that β_1 is unity. That is essentially the same as dividing by β_1, which would not affect any of the coefficients except the constant term, whose value is generally unimportant.

If one wanted to impose the additional restriction that the weights were zero at the far end, that would simply entail adding an additional term in (4.21) that said $w_5 = 0$. A similar constraint $w_0 = 0$ would be used if the weights were constrained to zero at the near end.

It may not be intuitively obvious what is accomplished by this method of calculating new variables with different coefficients to various powers. A formal exposition involves Lagrangian multipliers and more algebra than is appropriate here. However, on a heuristic level, we can think of fitting a quadratic (or a cubic) as an approximation of a more complicated lag structure that exists in the real world. Such an exercise is little more than curve fitting; but as pointed out earlier, theory doesn't tell us the length or the shape of the actual lag structure, even if it does suggest the variables and the approximate magnitudes of the coefficients that should be expected.

The PDL method is useful because it sharply reduces the number of degrees of freedom that are used in the estimation of the equation, it reduces multicollinearity, and it reduces the probability that one or two outliers will determine the shape of the estimated lag distribution. The major drawback to this method, as will be seen, is that for variables with strong trends, it is often difficult to determine empirically how long the lag should be, and what order the polynomial should be.

Two of the most common examples of the use of PDLs in macroeconomics are the consumption function, and the lag between capital appropriations and actual capital expenditures. In the case of the consumption function, Friedman estimated a 17-year lag on income to approximate permanent income, although his original work was done before the concept of PDLs were used in econometrics. The regressions of capital spending on appropriations by Almon was the seminal use of PDLs in econometrics.

While Friedman was correct in his belief that consumer spending depends on average or expected income, his lag of 17 years was far too long for two reasons. First, the average consumer does not have that long a memory. Second, consumers also look ahead to what they expect their income to be; the work on rational expectations had not yet been developed when Friedman published his pioneering *Theory of the Consumption Function*.[3] As a result, consumer spending generally does not depend on income with more than a three-year lag, and the use of PDLs is not usually utilized today to estimate consumption functions. The use of PDLs in the investment function is explored in case study 6.

4.3.3 General Guidelines for Using PDLs

The following guidelines should be useful when using PDLs.

1 Some textbooks suggest starting with a high-order polynomial and then dropping the insignificant terms; but in real life it does not make much sense to experiment with any polynomial higher than a cubic unless one has independent information that would suggest a more complicated lag structure.
2 Don't expect to be able to pinpoint the precise length of the lag structure. The R^2 will vary hardly at all for adjacent lags (e.g., 12 compared with 13).[4]
3 In long lags, the coefficients often have the wrong sign for a few months or quarters, which doesn't make any economic sense. However, it is often difficult to get rid of these erroneous signs without compromising the rest of the equation. My advice would be if the coefficients with the wrong sign are tiny, leave them in. "Tiny" in this context means a t-ratio with an absolute value of less than 0.5.
4 Often, the tendency is to keep adding lags as long as $\bar{R}^2$ keeps increasing. However, that generally tends to make the lag longer than is the case in the underlying population. As pointed out, the adjusted R^2 will keep increasing as long as the t-ratios of each additional term are greater than 1. However, in the case of PDL, this rule of thumb should be modified so that additional lag terms are not added unless the t-ratios of those additional terms exceed 2.
5 Since the average length of business cycles used to be 4–5 years, it is sometimes the case, when using time-series data before the 1980s, that long lags may simply be picking up a spurious correlation with the previous cycle; this is known as the echo effect. Since business cycles happen less frequently these days, splitting the sample period and reestimating the equations starting in 1982 should reduce if not eliminate any such effect. If the maximum length of lag is reduced in the more recent sample period, then the earlier correlation probably does represent an echo effect.
6 The use of PDLs is usually quite sensitive to the choice of variables in the equation. That makes testing the equation more complex. In general, you should include what are expected to be the relevant variables in the equation, rather

[3] Friedman, Milton, *A Theory of the Consumption Function* (Princeton University Press for NBER, Princeton, NJ), 1957.

[4] Further discussion on choosing the length of lag is found in Frost, P. A., "Some Properties of the Almon Lag Technique When One Searches for Degree of Polynomial and Lag," *Journal of the American Statistical Association*, 70 (1975), 606–12.

than trying to work out the precise form of the PDL on the key variable first and then watch it fall apart when other variables are added. Also, the equation is likely to be unstable if one uses several PDL variables with long lags.

Since several doubts have been raised about the efficacy of PDLs, what are some alternative suggestions?

(a) If weights of the distribution become insignificant after a short lag, estimate those terms directly and ignore the long tail. That is what we have done with the GDP term in the investment equation.
(b) Alternatively, if a function appears to have a peak with a fairly short lag but a long tail, try estimating the tail as a separate variable.
(c) If the mean lag seems to keep increasing indefinitely as more terms are added, set up another workfile with annual data and estimate the function with annual data and no PDL, and see if the result is about the same. If it is significantly different, the PDL method is picking up some spurious correlation.
(d) Finally, try calculating the equation in percentage change form, with each lagged variable entered separately. There will be a higher ratio of "noise" in the data points, but the parameter estimates will provide a clearer hint about whether the long lag structure is justified. Bear in mind, though, that the percentage change formulation usually *underestimates* the length of the underlying lag structure and the long-term elasticities.

4.4 Treatment of Outliers and Issues of Data Adequacy

Before actually estimating a regression equation, it is always – without exception – best to plot the data and examine their characteristics. Besides spotting any possible errors of transposition or copying that might have crept into the data, all of the statistical formulas and goodness-of-fit statistics are based on the assumption that the residuals are normally distributed. Most of the time, that is a reasonable assumption under which to proceed. However, if some of the residuals are exceptionally large, the least squares algorithm will overweight those observations, possibly distorting the parameter estimates, which would lead to inaccurate forecasts. This section discusses various methods for treating this problem.

4.4.1 Outliers

Suppose that after estimating what appears be a satisfactory regression equation, some of the errors for individual observations are more than three standard deviations from the mean. Assuming normality, the probability that they are drawn from the same population is less than 1 in 1,000. Generally it should not be assumed that these points are drawn from the same data generation function. The major options facing the econometrician are the following:

- remove those observations from the sample period entirely
- try to find economic variables that explain these aberrant points
- add specific dummy variables for those periods.

It is possible that these outliers are "harmless," which means the parameter estimates of the dependent variables will be about the same whether the outliers receive special treatment or not. In that case, adding a dummy variable simply improves the sample period fit without improving forecast accuracy. In such cases, the treatment of these outliers is irrelevant, and they can be ignored. However, that is usually not the case. It is more likely that the disturbances causing these outlying values also affect the independent variables; several examples are provided next.

Before turning to economic relationships, we look at the statistical distortion that can occur from a purely random observation by choosing an artificial example where bad data are introduced into the sample observations. To illustrate how outliers and dummy variables can affect an equation, consider the equation in which the percentage change in consumption is a function of the percentage change in income and the yield spread with an average lag of half a year. The sample data are then altered by introducing one period of "bad" data for both consumption and income, and the regressions are recalculated with and without dummy variables to offset these outlying observations. The bad data used here are constructed by adding 100% to the actual change; e.g., if income rose 5% in that year, the bad data would show an increase of 105%. While such large errors are not likely to be encountered in actual forecasting, exaggerating the case emphasizes the distortions that can occur in the parameter estimates. The results are summarized in table 4.2.

It is clear that if the outlying data points are ignored, the equation is ruined. Not only is the R^2 close to zero and the coefficients insignificant, but their values are far different and, in one case, the value of the yield spread variable switches signs. If the error occurs in the *independent* variable and a dummy variable is added, the equation in this case is unchanged. If the error occurs in the *dependent* variable and a dummy variable is added, the resulting R^2 is far overstated, although the parameter estimates do not change. With an error of this magnitude, though, it is clear that some adjustment must be made: either a dummy variable must be added or the erroneous observations must be removed from the sample.

If dummy variables are used, it is usually a good idea to recalculate the regression without those data points as a cross-check to make sure the equation has not changed very much. That will also provide a better estimate of the underlying value of R^2 for this regression.

If the outlier occurs in the dependent variable, but that value is uncorrelated with any of the independent variables, regressions that are calculated (i)

Table 4.2 Illustration of statistical distortion

Equation	$\bar{R}^2$	SE	DW
$\%C = 0.35 + 0.765{*}\%Y + 0.32{*}\text{YLD}$ (8.7) (3.0)	0.734	0.98	2.10
$\%C\# = 1.69 + 1.688{*}\%Y - 0.77{*}\text{YLD}$ (1.1) (0.4)	–0.022	1.71	2.09
$\%C\# = 0.37 + 0.774{*}\%Y + 0.31{*}YLD + 99.1{*}\delta_C$ (8.8) (2.9)	0.997	0.98	1.95
$\%C = 0.26 + 0.008{*}\%Y\# + 0.48{*}YLD$ (0.4) (2.4)	0.105	1.79	0.97
$\%C = 0.35 + 0.765{*}\%Y\# + 0.33{*}YLD - 76.5{*}\delta_y$ (8.6) (2.9)	0.725	0.99	2.10

$\%C$ = percentage change in consumption; $\%C\#$ = same as $\%C$ with 100% added to one observation; $\%Y$ = percentage change in disposable income; $\%Y\#$ = same as $\%Y$ with 100% added to one (different) observation; YLD = yield spread, lagged half a year. δ are dummy variables for C and Y respectively (1 in the outlying period, 0 otherwise). Numbers underneath the coefficients are t-ratios.

excluding the outlier, and (ii) including an outlier with a dummy variable for that period, will yield approximately the same coefficients, t-ratios, and standard error of estimate. The only difference is that R^2 will be much higher for the second equation.

Suppose the outlier is found in the independent variable. Then the coefficient of that variable is the same (i) without the bad data, and (ii) with the combination of the bad data and the dummy variable if in fact the cause of the outlying observation is uncorrelated with the independent variables. Had there been some correlation, the value of the Y coefficient would have differed from the equation with no outlier and no dummy variable.

Thus the appropriate test to be performed is to run the regression with and without the outliers (using a dummy variable when including the outlier) and see how much the coefficients change. If they are about the same, and the t-ratios are significant both with and without the outliers, the general rule of thumb is not to worry about these outliers; while they can be handled with a dummy variable, they will not affect forecasting accuracy.

The more serious issue occurs when the outlying observations are correlated with some of the independent variables. Consider, for example, a widespread reduction in production – due either to a natural disaster or a strike – that reduces sales and personal income because some workers are not being paid. Shortages would ordinarily be accompanied by higher prices; but if there is rationing, the price equation would not work properly and a dummy variable would be appropriate. Other cases are discussed in more detail in the next section. In a related example, suppose a major electric power plant fails and customers in that service area must purchase power from a nationwide grid at five times the usual price – a result that can occur under deregulation. That sample point is clearly drawn from a different underlying data generation function, yet there is a significant correlation between quantity and price. Simply treating that observation with a dummy variable may distort estimates of the actual price elasticity.

Before turning to the actual use of dummy variables in equations, however, it is useful to discuss briefly some of the problems that can occur with data series. These problems may or may not necessitate the use of dummy variables, but data errors will generally distort the parameter estimates just as much as actual outlying events. Hence the use of a dummy variable may be merited if the residual is well outside the sample period range, even if there does not appear to be any real-world event that would explain the outlying observation.

4.4.2 Missing Observations

Suppose one is estimating a regression equation and some of the data are missing. What is the best way to proceed?

One possibility is simply to omit all those observations from the sample period. However, that is not always advisable. Suppose the independent variable in question is used in a 20-quarter distributed lag; then 20 observations would have to be omitted for each missing data point. Sometimes data are available in slightly different form and can be combined – spliced together. In other cases, quarterly data can be interpolated from annual series, or monthly data from quarterly series.

Naturally there is some risk in making up the data with a certain hypothesis already in mind, and then finding that the data support that hypothesis. All researchers would always prefer to have complete data sets. But when that is not possible, what are the "second best" alternatives?

The problem with omitting observations from the sample period is intensified when long lag structures are used. Suppose one of the independent variables enters the equation with a distributed lag of 20 quarters, and the entire sample period is only 80 quarters. The sample period has already been reduced from 80 to 60 observations to accommodate this lag. To lose another 20 observations just because of one single missing data point could reduce the sample size to the range where the results are inconclusive.

Another case arises with the equation for purchases of motor vehicles. Monthly and quarterly data are available for auto sales starting in 1959, domestic light truck sales starting in 1966, and foreign light truck sales starting in 1976. Before that, only annual data are available. One option is to start the equation in 1976; however, that reduces the sample period almost by half. Since light truck sales, and particularly foreign light truck sales, were not very important in the early years (presumably one of the reasons that monthly and quarterly data were not prepared), the results might be improved by extrapolating those data for the earlier period. The various ways of filling in missing data points include the following.

1 Simple interpolation is acceptable if the missing point appears in a series with a strong trend and a small variance around that trend.
2 If there is no trend and the observations appear to be serially uncorrelated, one could simply use the mean value of the variable. That might be the case for percentage changes, for example. If one is taking deviations around the trend, the missing value would then be 0.
3 Suppose one series (e.g., foreign truck sales) are available only on an annual basis for part of the sample period, while a closely related series (domestic truck sales) are available monthly or quarterly. Then one can interpolate monthly or quarterly series for foreign truck sales based on the annual data for that series and monthly and quarterly figures for domestic sales.
4 Calculate a regression relating the data series with the missing observation to other variables during the sub-sample period when all the data are available; then use the "predicted" value for missing observations. That is equivalent to estimating the equation by omitting sample points where data are missing if none of the variables is lagged; but where lagged values are used, the size of the sample period can be expanded.
5 In these examples it is assumed that monthly and quarterly data are seasonally adjusted. If they are not, the appropriate seasonal factor should be added to that month or quarter when estimating the missing data.

To a certain extent, it is not known how well these methods work, because by definition the missing data do not exist (although experiments can be constructed where one "pretends" not to have some of the observations). In general, though, (3) and (4) usually work fairly well where the relationships fit well during the periods when all the data are available. Conversely, assuming the value is equal to some sample period average generally does not work well and should be tried only as a last resort.

4.4.3 General Comments on Data Adequacy

It is a truism, yet one that can be repeated often, that the estimated model is only as good as the underlying data. While examples of inaccurate data abound, it is useful to group them into the following general classifications:

- *Outright errors.* These are input errors, or changes in definitions that are not properly reflected in the published data.
- *Data revisions.* This is often a serious problem for macroeconomic data. The corrections are usually due either to (i) missing data in preliminary releases, or (ii) change in sample survey techniques. In some cases, such as the inclusion of

business purchases of software in capital spending, the entire concept of the term is changed.

- *Restatement of profits or other company information.* To a certain extent, some of this is due to mergers, acquisitions, or divestitures, but the main problem is retroactive writedowns.
- *Fraud.* In the international arena, intentional fraud may occur when the government wants to make the country's output appear better than was actually the case (primarily, but far from exclusively, the case in former communist regimes).
- *Defects in survey method or obsolete surveys.* For example, the CPI weights could be based on the market basket consumers bought almost a decade ago rather than what they buy today. If one is focusing primarily on the prices of apples and oranges, it probably doesn't matter. If the area of interest is CD-ROMs and Internet access, it probably does.
- *Lack of understanding of how to collect the underlying data.* This is probably more often the case for LDCs, although occasionally a new series even for the US will have to be completely revised when the underlying process is understood more thoroughly.
- *Changes in growth patterns due to rebasing.* In most cases, the growth rate will be reduced by moving to a more recent base year. However, in the case of computers, where the deflator declines, the opposite is true.
- *Seasonal data.* Seasonal patterns do change over time, but no method of adjusting for seasonal data is perfect, and sometimes these methods distort the underlying data. Also, when seasonal factors are revised, monthly or quarterly changes are often far different in the revised data.
- *Reclassifications.* Reclassification of companies from one industry to another often distorts industry data. This is particularly severe in the case of conglomerates, where small percentage change may shift the entire company from industry A to industry B.
- *Consumer misreporting.* For individual consumer data, individual income, assets, spending patterns, or saving might be misstated or misreported. In particular, people might understate their income because some of it was not reported to the Inland Revenue.

There are no "textbook" answers about how to determine whether the desired data series are reliable, and no exhaustive list that will include all possible data deficiencies. The above list, however, covers most of the areas where data problems occur.

4.5 Uses and Misuses of Dummy Variables

In general, a dummy variable takes a value of 1 during designated periods and 0 elsewhere. For a series with constant seasonal factors, the seasonal dummies for quarterly data are 1 in the *i*th month or quarter and 0 otherwise. Sometimes, however, dummy variables either take on a variety of values, or they are combined with other terms. Thus in addition to seasonal dummies, one can distinguish among the following principal types of dummy variables. Examples are provided for each of these cases.[5]

- single or isolated event changes: wars, energy crises, strikes, weather aberrations
- changes in institutional structure: floating of the dollar, deregulation of the banking sector
- changes in slope coefficients: variable becomes more or less important over time
- nonlinear effects: big changes are proportionately more important than small changes
- ceilings and floors: economic variables have a larger impact above or below a certain level.

As already noted, it is a simple matter to boost the sample-period goodness-of-fit statistics with dummy variables without improving forecasting accuracy. This section considers some of the economic issues.

Dummy variables are often properly introduced to reflect institutional changes: deregulation of a particular sector or industry. Compare, for example, the behavior of the airline, trucking, banking, or stock market sectors before and after deregulation.

Another type of far-reaching change could occur because of changed expectations. For example, when the Fed did not have a credible monetary policy, declines in the unemployment rate were widely thought to presage higher inflation. However, once credibility was reestablished, the tradeoff between unemployment and inflation disappeared.

Company data for sales, orders, profits, etc., would obviously change if the company acquired another entity, or divested some of its divisions. For accounting and investment purposes, earnings per share can be restated so they are comparable, but any time-series data for total sales would show major shifts.

In terms of econometric application, one of the primary issues is whether the dummy variable should be applied to the constant term of the equation, to some of the slope coefficients, or possibly to the entire equation. In the latter case, the same functional form can be reestimated for two or more subperiods.

It was already shown in the previous section that erroneous data can, under extreme circumstances, seriously distort the parameter estimates. That case was exaggerated to emphasize the point, which is that the significant criterion is whether the dummy variable is correlated with the other independent variables. If the outlier is due to a truly random event, then omitting a dummy variable will reduce the sample period fit but leave the parameter estimates unchanged. However, if the dummy variable is correlated with the other independent variables, then omitting it will bias the other coefficients.

[5] For further discussion see Pindyck and Rubinfeld, pp. 122–5; and Johnston and DiNardo, pp. 133–9.

4.5.1 Single-event Dummy Variables

Consider the case of an auto strike. Consumers buy fewer cars because they realize there will be fewer choices in the showroom, so they may not be able to find their preferred make and model, and may also receive a smaller discount. If the strike is lengthy, that will not only affect the auto industry but the economy in general: disposable income will fall and, although striking workers are not counted as unemployed, the unemployment rate will rise as other workers are laid off. Without the use of a dummy variable for strikes, the coefficients for income and the unemployment rate would probably be overstated.

However, that is not the end of the story. The auto strike delayed sales, but it probably did not cancel them. The loss of sales during the strike period is generally made up in the following period. Thus the appropriate value of the dummy variable would be –1 during the strike and +1 in the next period. In other industries, such as the steel industry, if the strike were anticipated, the dummy variable might be +1 in the period before the strike and –1 during the strike. If the strike lasted longer than expected, the dummy variable might have the values +1, –2, and +1. Only if there were a permanent loss of sales would the values of the dummy variable sum to less than zero.

The same argument can be made for dock strikes: exports and imports both rise before the strike occurs, decline during that period, surge the next period, and then return to trend levels. In that case, the values of the dummy variable would also sum to zero over the period before, during, and after the strike.

In some cases, the sum of the values of the dummy variable might be greater than zero. Suppose a hurricane devastates coastal areas. In the period before the hurricane, construction is at normal levels. After the storm ends, construction expenditures rise sharply for a while. If the rebuilding phase lasts several periods but gradually tapers off, the dummy variable might take the values 4, 3, 2, and 1 in the four periods following the storm. On balance, though, construction activity over the entire period will be higher than if no storm had occurred.

Sometimes the interaction is more complicated. The Nixon Administration imposed a wage and price freeze on August 15, 1971, that lasted for 90 days. That was followed by Phase II of controls, lasting through the end of 1972, during which wages and prices could rise by only a certain percentage determined by the government. During Phase III, which started on January 1, 1973, prices could be raised only by the amount that costs increased. Controls were ended on May 1, 1974, at which point prices briefly rose by record amounts.

It might seem clear that a dummy variable that was negative during controls and then positive for a while would be appropriate. In some commodity price equations the dummy variable is important, as will be shown later. However, when the macroeconomic inflation rate is correlated as a function of labor costs, money supply, and oil prices, the residuals do not show any such pattern. A wide variety of dummy variables this author tried are not significant.

In this case the reason is not obvious. The inflation rate is, and should be, negatively correlated with productivity growth. During the period of controls, many firms had an incentive to understate the rise in prices and hence overstate the rise in output, since the current-dollar numbers could not be easily fudged. As a result, reported productivity growth soared to 3.5% during the period of controls and then declined to –1.6% when they were removed. It is quite unlikely that such a pattern actually occurred.

Yet both theory and empirical evidence suggest that productivity rises faster when inflation is lower, because capital goods are then purchased in order to earn their highest real rate of return, rather than being purchased as a hedge against inflation. Hence the strong negative correlation between productivity and inflation is theoretically as well as empirically robust. However, because of faulty data during that period, the correlation may be overstated. The use of a dummy variable should reduce that parameter estimate – if we had accurate data; but it is not available. In this particular example, a dummy variable is theoretically reasonable, but is not empirically significant.

This is perhaps an extreme example, but it illustrates how the use of dummy variables depends in large part on the correlation between the dummy variable and other independent variables. When they are correlated, it is good statistical practice to include a dummy variable: in that case, when used within reason, it is not just merely curve fitting or ad hoc adjustment.

4.5.2 Changes in Dummy Variables for Institutional Structure

Over the past 25 years or so the US economy has undergone many structural changes involving deregulation. One of the most important was the deregulation of the banking sector in the early 1980s. Before then, growth in the money supply (M2) closely followed changes in the monetary base – required reserves plus currency – which could be closely controlled by the Fed. Since then, there has been no correlation between percentage changes in the monetary base and M2. Thus in estimating an equation for the percentage changes in M2, it is best to multiply the percentage changes in the monetary base by DBR – a dummy variable for changes in banking regulations – which is 0 through 1980.3 and 0 afterward. A comparison of these two series shows that, starting in late 1980, money supply growth accelerated at the same time that the growth rate in the monetary base decreased.

After deregulation, changes in the money supply were more closely correlated to loan demand than to the monetary base, so the changes in business loans are multiplied by $(1 - DBR)$. Also, the spread between the Federal funds and the discount rate, while still a significant determinant of changes in the money supply, is much less important after 1980 than before, so that variable is also multiplied by *DBR*. After 1980 the term still has a negative sign but is only marginally significant. The dummy variable is also estimated as a separate term, otherwise there would be a discontinuity when the contribution of the monetary base dropped to zero.

The estimated equation is as follows. The first number after the symbol is the length of lag; most variables are four-quarter percentage changes. The number after the symbol in parentheses is the period when the lag starts.

$$
\begin{aligned}
&\%M2,4\\
&\quad = \underset{}{2.86} + \underset{(21.5)}{0.991 * \%MBASE,4 * DBR} - \underset{(7.0)}{1.02 * (FFED(-1) - DISR(-1)) * DBR}\\
&\quad\; - \underset{(2.6)}{0.91 * DBR} - \underset{(8.7)}{0.51 * \Delta INFL(-1),4} - \underset{(3.2)}{0.16 * \Delta INFL(-5),4}\\
&\quad\; + \underset{(17.3)}{0.152 * \%LOAN(-4),12 * (1 - DBR)}\\
&\qquad\qquad\qquad RSQ = 0.853;\ DW = 0.60.
\end{aligned}
\tag{4.24}
$$

In this equation *M2* is the money supply, *MBASE* is the monetary base, *FFED* and *DISR* are the Federal funds and discount rate, *INFL* is the rate of inflation, and *LOAN* is commercial and industrial loans.

4.5.3 Changes in Slope Coefficients

Sometimes, variables become more or less important over the sample period because of changes in institutional structure. For example, before 1985, an increase in nominal income pushed many taxpayers into a higher tax bracket even if their real income did not rise. For that reason, Federal tax receipts were highly correlated with inflation. After 1985, however, the tax brackets were indexed, which means the upper end of each bracket increased by the percentage that the CPI rose the previous year. Nominal increases that were not accompanied by real increases no longer caused taxpayers to move into a higher bracket. Hence the equation for personal income tax receipts shows that inflation and wage rates are much more important before 1985 than afterwards. These terms are thus multiplied by a dummy variable that is 1 before 1985 and 0 afterward.

Another example might occur in situations where foreign trade has become an increasingly large proportion of total sales. In previous periods, changes in the value of the dollar would have relatively little impact on sales, but in recent years that variable has become increasingly important. In that case, the value of the dollar would be multiplied by a time trend.

4.6 Nonlinear Regressions

So far we have avoided nonlinear regressions for a very simple reason. All the standard statistical tests discussed in this book are based on the assumption that the regressions are linear in the parameters. If they are not, these tests will not report the correct levels of significance. As a result, virtually all estimation procedures for econometric equations where nonlinear relationships may occur are based on linear approximations to nonlinear equations. This section discusses the most common methods of linearizing equations, which can be grouped into the following categories:

- log-linear
- quadratic and other powers, including inverse
- ceilings, floors, and Kronecker deltas
- piecewise linear.

4.6.1 Log-linear Equations

The most common case of an equation that is nonlinear in the variables is one that is linear in its logarithms. As already noted, logarithms would be appropriate if the elasticities were constant throughout the range of the independent variables. This assumption is often made in the case of demand and supply functions. Two standard cases are considered: the demand for gasoline by consumers, and the aggregate production function.

Consumer demand for gasoline is identically equal to the number of registered motor vehicles times the amount each vehicle is driven each year times the average fuel economy per vehicle (miles/gallon). This relationship can be written as:

$$CGAS \equiv \text{number of vehicles} * \text{miles/year} \div \text{fuel economy}. \tag{4.25}$$

In the short run, an increase in the price of gasoline might result in fewer miles driven per year, whereas in the long run it might result in the purchase of more fuel-efficient motor vehicles. Also, a rise in the price of gasoline would reduce real disposable income, which might reduce the demand for new motor vehicles, although any such relationship would be captured by the income term.

Elsewhere in this text an equation is presented to explain the purchase of new motor vehicles. That is obviously not the same as the total number of vehicles on the road, which is a function of long-term trends in income and demographics and can be represented by those variables. The number of miles driven per year is a function of income and *short-term* changes in the relative price of gasoline, while average fuel economy depends on *long-term* changes in the relative price of gasoline.

First we show that real consumption of gasoline can be explained by the number of motor vehicles and average miles per gallon (MPG); the number of miles/year turns out not to be significant. *MPG* is then a function of a 12-year lag on the relative price of gasoline, reflecting the fact that most changes occur when consumers save by buying more fuel-efficient cars, not by reducing the number of miles traveled.

Consumption of gasoline in constant prices is thus a function of the number of motor vehicles registered and a 12-year average of the relative price of gasoline; the form used is PDL (12,3,2). Because of significant positive autocorrelation, the function is reestimated with an AR(1) term. The same equation is then estimated in logarithmic terms, both with and without an AR(1) term. Finally, these equations are compared with a percentage change equation using the same terms and lag structure.

All these functions look about the same in terms of goodness-of-fit statistics and patterns of residuals, but the elasticity estimates are significantly different. The price elasticities for gasoline in each form of the equation are:

Levels, no AR	–0.24
Levels with AR(1)	–0.42
Logs, no AR	–0.26
Logs with AR(1)	–0.39
Percentage change	–0.39

The equations in levels and logs yield about the same estimates of the elasticity without the AR term, but they are both much lower than with the AR term. The latter appears to be more accurate, since that is also the value obtained when percentage changes are used and autocorrelation is not present. The long-term price elasticity of gasoline for consumer use is thus estimated to be –0.4. Note that in this case, the choice of levels or logs makes little difference, but adjusting the equation for autocorrelation makes quite a bit of difference.

The next equation considered is a macroeconomic production function: constant-dollar GDP is a function of labor input, capital input, and the growth in technology, which is usually represented with a time trend. Such functions have been standard ever since the concept was introduced by C. W. Cobb and Paul Douglas in 1928.[6] The theoretical development assumes that a given percentage change in both labor and capital inputs causes in the same percentage change in output, which means the equation is linear in the logarithms. In this case all variables have strong trends, so the issue of mixing variables with and without trends does not arise.

[6] Cobb., C. W., and P. H. Douglas, "A Theory of Production," *American Economic Review*, 18 (1928), 139–65; and Douglas, P. H., "Are There Laws of Production?" *American Economic Review*, 38 (1948), 1–41.

According to the national income statistics, labor income accounts for about two-thirds of GDP and capital income for about one-third, so those ought to be the coefficients in the logarithmic equation. In fact the equation is

$$\log GDP = -0.60 + 0.626 * \log L + 0.411 * \log K(-1) + 0.0066 * t$$
$$(4.9) \qquad\qquad (4.6) \qquad\qquad\qquad (2.1)$$
$$RSQ = 0.998;\ DW = 0.48. \tag{4.26}$$

The linear equation is far inferior; the results are

$$GDP = -1491 + 0.0384 * L + 66.37 * K(-1) - 50.65 * t$$
$$(3.0) \qquad\qquad (5.0) \qquad\qquad\qquad (4.8)$$
$$RSQ = 0.996;\ DW = 0.25. \tag{4.27}$$

In both these equations, *GDP* is in constant dollars, *L* is non-farm payroll employment, and *K* is capital stock.

The results from the levels equation are considered far inferior because if the elasticities are calculated at the mean, the elasticity for labor is 0.70, which is quite reasonable, but the elasticity for the capital stock is 0.96. Also, the time trend has the wrong sign. In this case, the log equation is clearly superior.

However, the log estimates also leave something to be desired. The coefficient of the time trend is only 0.0066, implying an average annual advance in technology of 0.66%; the figure is usually estimated to be between 1% and $1\frac{1}{2}$%. Also, the very low *DW* statistic can be worrisome; that led to an incorrect estimate of price elasticity in the consumption of gasoline equation.

Since the variable on the left-hand side is actual as opposed to potential GDP, the variables on the right-hand side of the equation ought to be utilized labor, which is employment times hours worked, and utilized capital stock, which is total capital stock times capacity utilization. If these variables are substituted, the results are

$$\log GDP = -4.27 + 0.667 * \log LH + 0.268 * \log KU + 0.0129 * t$$
$$(5.6) \qquad\qquad\qquad (5.4) \qquad\qquad\qquad (6.1)$$
$$RSQ = 0.998;\ DW = 1.10 \tag{4.28}$$

and

$$GDP = -3336 + 0.00257 * LH + 0.1252 * KU - 12.61 * t$$
$$(8.0) \qquad\qquad\qquad (1.2) \qquad\qquad\qquad (1.3)$$
$$RSQ = 0.995;\ DW = 0.47 \tag{4.29}$$

where *LH* is employment times hours worked per week and *KU* is capital stock times the rate of capacity utilization in the manufacturing sector.

The results are substantially improved in the log equation. The coefficient of the labor term is exactly 2/3, and the coefficient of the capital stock term is not significantly different from 1/3, although it is a little low. The coefficient of the time trend is equivalent to 1.3% per year, almost precisely in the middle of the 1% to 1½% range calculated by economists who measure productivity separately. The levels equation has not improved at all, and the time trend still has the wrong sign. In this case, using a logarithmic function clearly improves the results.

4.6.2 Quadratic and Other Powers, Including Inverse

We referred briefly to the concept that including variables with strong trends and no trends in the same equation may lead to forecasting errors. By the same token, estimating an equation in which Y is a function of both X and X^2 is also likely to lead to additional forecasting errors. If Y and X are linearly related, the use of X^2, while it may improve the sample period fit, will almost always increase forecast error as that variable continues to grow more rapidly. Hence terms with powers different from unity should generally be used with trendless series. That usually means using actual or percentage first differences to powers other than unity, including the inverse of the variable. Here too one must utilize a certain degree of caution; squaring the first difference of series that contained both positive and negative numbers would result in all positive numbers, which presumably is not what is desired.

From an economic perspective, one likely use of nonlinear powers would occur in circumstances where big changes are proportionately more important than small changes. If the price of newsprint rises 5%, publishers may grumble about it but they are not likely to make any changes. However, if the price were to rise 50% they might invest in new technology that would make newspaper pages thinner. If an employee receives a 3% wage hike in real terms, he is not likely to change his lifestyle and buy a more expensive home, but if he receives a 100% increase because of a promotion or a new job, a move is much likelier.

An example can also be drawn from the agricultural sector. In most years, the price of soybeans is closely related to the price of corn and the price of livestock. However, in years when shortages are likely to occur, prices rise much more rapidly. Hence the price is related to the inverse of soybean stocks. This variable is also multiplied by *@SEAS(6)*, a dummy variable that is 1 in June and 0 elsewhere, because that is the month when prices rise the most during years when shortages are likely. That variable is highly significant, whereas the reader may verify that the same term is insignificant in a linearized version of the same equation. In this equation, *SOYSTOCK* is the stock of soybeans, *MOVAV* stands for moving average, and all other variables are prices received by farmers:

$$\begin{aligned}
&\mathit{SOYBEAN\ PRICES}\\
&\quad = -0.017 + \underset{(13.5)}{1.901} * \mathit{CORN\ PRICES} + \underset{(3.4)}{0.852} * \mathit{LIVESTOCK\ PRICES}\\
&\quad + \underset{(4.1)}{37.1} * @\mathit{SEAS}(6) / @\mathit{MOVAV}(\mathit{SOYSTOCK}(-1), 12) + \underset{(42.0)}{0.921} * \mathit{AR}(1)\\
&\qquad \mathit{RSQ} = 0.933;\ \mathit{DW} = 1.80
\end{aligned} \tag{4.30}$$

4.6.3 Ceiling, Floors, and Kronecker Deltas: Linearizing with Dummy Variables

In this type of formulation, a variable has little or no effect up to a certain level, then the relationship becomes increasingly important. For example, suppose a firm has excess capacity. If sales increase, initially there will be no need to expand. After a while, though, further increases in sales would result in more net capital spending. Suppose that a capacity utilization rate of 80% represents an average crossover point for the overall economy. Then we could write

$I_{net} = b * \delta\ (CP - 80)$,

where CP is the rate of capacity utilization, and δ has a value of 1 when $CP >$ 80 and 0 otherwise. Such a term is known as a Kronecker delta.

This formulation works quite well when used to explain purchases of industrial equipment. The two independent variables are the Aaa corporate bond rate, lagged 1–5 years, and the rate of capacity utilization in the Kronecker delta form, lagged from 2 to 10 years. The reader can verify that using the overall rate of capacity utilization without the dummy variable provides substantially inferior results (*SE* of 0.00089 compared with 0.00069). The fit of the equation is shown in figure 4.1. The equation, estimated with annual data from 1959 through 1999, is

$$\begin{aligned}
&\mathit{INDEQP}/\mathit{GDP}\\
&\quad = 0.0174 - \underset{(7.6)}{0.000621} * \mathit{RAAA}(-1) + \underset{(3.3)}{0.000166} * \mathit{DCU}80\\
&\quad + \underset{(9.5)}{0.000483} * \mathit{DCU}80(-1) + \underset{(10.1)}{0.00067} * \mathit{LAGDCU}80\\
&\qquad \mathit{RSQ} = 0.931;\ \mathit{DW} = 1.02
\end{aligned} \tag{4.31}$$

Where *INDEQP* is purchases of industrial equipment in constant dollars, *RAAA* is the real Aaa corporate bond rate (nominal rate minus inflation rate) lagged 1 to 5 years, $DCU80 = \delta(CP - 80)$ as described above, and *LAGDCU80* is a PDL starting with lag 2, extending back for eight more years, set to 0 at each end.

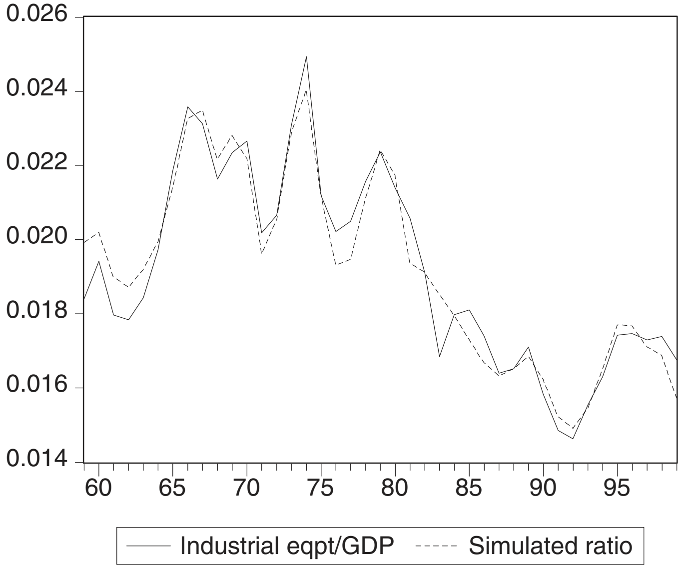

Figure 4.1 Ratio of purchases of industrial equipment to GDP, both in constant dollars, as a function of the Aaa bond rate and lagged rate of capacity utilization when over 80% (term is zero otherwise).

Microeconomics theory suggests that diminishing marginal returns results in rising costs and prices in the short run when the rate of capacity utilization rises above normal levels. That formulation would imply the use of a Kronecker delta in the equation for the producer price index. However, the term is not highly significant, suggesting that in most cases firms raise prices when costs increase but do not boost margins when capacity utilization rates are high. The equation is

$$\begin{aligned}
&\%PPI \\
&\quad = 0.30 + \underset{(17.6)}{0.110} * \%POIL + \underset{(10.0)}{4.36} * DWP - \underset{(4.3)}{0.100} * RFED(-2),4 \\
&\quad + \underset{(10.1)}{0.166} * \%ULC,4 + \underset{(2.8)}{0.086} * DCU80(-1),2 \\
&\qquad RSQ = 0.852;\ DW = 1.36
\end{aligned} \tag{4.32}$$

in which *PPI* is the producer price index for industrial commodities, *POIL* is the *PPI* for oil prices, *DWP* is a dummy variable for wage and price controls during the Nixon Administration, *RFED* is the real Federal funds rate, *ULC* is unit labor costs, and *DCU80* is the Kronecker dummy variable described above.

4.7 General Steps for Formulating a Multivariate Regression Equation

We have now covered some of the most common pitfalls found in estimating single-equation time-series structural forecasting models; chapter 5 presents the most common tests of structural stability, Part III examines time-series regressions – i.e., non-structural models – and Part IV covers a variety of forecasting models. Hence this is an appropriate point to summarize the key rules to keep in mind when building single-equation structural forecasting models. The following general checklist indicates the usual steps that should be followed.

Step 1
Determine the underlying economic theory.

Step 2
Collect data and check for possible glitches in the data, sudden shifts in patterns, or outliers that need further explanation. Always plot the data series before actually calculating any regressions.

Step 3
Determine the lag structure. Often this will not be known in advance, but there should be some apriori structure to use as a starting point.

Step 4
Estimate the first pass at the regression. Most of the time, the results will be disappointing. Some variables will have the wrong sign and others will be insignificant. The residuals are likely to exhibit autocorrelation or heteroscedasticity. For all these reasons, further refinements are necessary. But it is useful at this juncture to consider whether the insignificant variables and wrong signs occurred because:

(a) the wrong theory was chosen
(b) empirical choice of theoretical variable is wrong
(c) lag structure is wrong.

It is always possible that (a) will turn out to be the answer, but before testing an alternative theory, (b) and (c) should be examined thoroughly. Search for other variables that will more closely represent the concept you are trying to measure. For example, someone might be trying to measure expected changes in, say, prices or interest rates. Since future events cannot be used in regression equations, try to figure out what past events best represent people's expectations of the future. It is not unreasonable to attempt several different formulations when trying to determine this.

Since theory does not indicate the precise form of the lag structure, it is appropriate to experiment with several different lag structures. However, it is generally inadvisable to use lags of 1, 2, 3 . . . , 10 in the same regression, because high multicollinearity will generally cause alternating signs. Some judgment is required. That is where PDLs can be useful.

Step 5

Virtually everyone who runs regressions compares their results by looking at R^2, including this author. Yet as noted previously, there is no point in improving the fit only to end up with the wrong signs or distorted coefficients, as well as all the other pitfalls listed above.

Step 6

If the variables seem to be measuring a common trend, consider other forms of the equation, such as percentage first differences, and see whether the results are consistent. That is particularly useful in time series with strong trends.

Step 7

Look at the residuals, and try to correlate them with some other variables or lag structures that haven't yet been tried; this can be done by graphing the residuals and the additional variable under consideration. Repeat the experiment until, at a minimum, all the variables are significant, the coefficients have the right sign, and the elasticities are reasonable.

Further steps

At this point, the researcher might think most of the work is done. Actually, for those building forecasting models, it is just starting. The estimated equation should now be put through a battery of tests designed to determine whether the goodness-of-fit statistics are really as high as the equation says, whether the coefficients are stable or unstable, whether they are biased or not, and whether this equation is likely to generate accurate forecasts. We will turn to these tests after looking at two case studies where distributed lags play an important role: the consumption function with quarterly data, and capital spending.

Case Study 5: The Consumption Function

Earlier in this chapter I discussed estimating the bivariate relationship between consumption and income in several different forms: levels, logs, first differences, and percentage changes. This case study shows what happens when the consumption/saving function is expanded to include monetary variables, expectational variables – inflation and unemployment – and other key economic variables, and examines how these variables change when the form of the equation varies.

According to the modern theory of the consumption function, consumption depends on some measure of average or expected income, not just current income. Expected income could be measured by some weighted average of past income, but variables representing consumer wealth – home prices and stock values – also measure expected income. Interest rates are an important determinant of consumption not so much because the cost of borrowing is important, but because lower interest rates also mean greater credit availability and the increased ability to refinance home mortgages at lower rates.

Attitudinal variables, such as the rate of inflation and the rate of unemployment, are also important; these are sometimes subsumed in an index of consumer attitudes, which is then entered as a separate variable. This approach is not used here because, as discussed later in this text, the fluctuations in consumer attitudes that are not related to inflation, unemployment, and stock prices do not appear to be correlated with consumer spending.

Hence the theoretical function says consumption is a function of current and lagged income, the Aaa corporate bond rate, the S&P 500 index of stock prices, the unemployment rate, the change in oil prices, and the relative price of homes. Given that function, we now consider the lag structure for each of these variables.

Since the changes in unemployment and inflation are expectational variables that affect the timing rather than the overall level of purchases, those lags should be relatively short. On the other hand, based on theoretical considerations, the lags for income, bond yield, stock prices, and home prices might be substantial. The first pass, then, is to use PDLs of 12 quarters, cubic polynomials, and constrained at the far end; 12,3,2 in the EViews formula.

However, these results are not useful. In particular, except for income, all the signs quickly reverse as the lags increase. Further experimentation, which the user may try, shows that even as the lag structure is shortened, the signs flip-flop. In the end, the optimal structure for the bond yield, stock price index, and relative price of homes has only a one-quarter lag. Even the coefficients of the income term drop off quickly, although they then recover with a longer lag. This indicates different lag structures for durable goods and services.

Having established a long lag on income and shorter lags on all other variables, we next consider whether these results are tainted by multicollinearity by using the battery of alternative formulations developed in this chapter: OLS, WLS, logs, percentage change, ratio, and deviations from logarithmic trends. Since this methodology has already been described in some detail earlier in this chapter, only the summary *t*-statistics are presented here (see table 4.3).

All equations except the one using percentage changes are adjusted with the AR(1) transformation. WLS is not listed since that is not applicable with the AR(1) transformation; without that adjustment, OLS and WLS were similar in all respects. In the ratio equation, the dependent variable is the ratio of consumption divided by disposable income; in that equation, the percentage change in income correctly has a negative sign, reflecting the fact that consumption adjusts to income with a lag.

These results are fairly instructive. Starting with the levels function, one could readily draw the conclusion that the stock market variable is much more important than the relative home price variable. That is also the case for the ratio function. In the logarithmic function, relative home prices are barely significant. In the percentage change equation, random changes overwhelm the

Table 4.3 Results for case study 5.

Form of equation	Income	*t*-ratios for:				
		Bond yield	Stock prices	Change in unemployment	Inflation	Relative home prices
Levels – OLS	68.6	−3.5	7.1	−2.5	−2.8	2.2
Logs	44.2	−8.4	3.4	−3.0	−2.4	1.8
Percentage change	4.5	−2.9	2.2	−3.7	−2.3	1.2
Ratio	−6.5[a]	−4.3	6.1	−1.2	−4.9	3.7
Deviation from trend	18.0	−2.3	2.5	−2.8	−2.3	3.1

[a] Percentage change in income over four quarters.

underlying function, and here too home prices are not significant. That might be expected because this is a longer-term effect. However, increasing the length of lag in the percentage change equation does not improve the fit, and the *t*-ratio for the change in stock prices turns negative.

In the deviations-from-trend equation, the results change significantly. While stock prices are still significant, the *t*-ratio drops sharply, and the importance of relative home prices improves sharply. An examination of the residuals from the levels and deviations-from-trend equations shows that, while both fail to track the negative impact of the first oil shock on consumption, the deviation-from-trend equation fits much better in the 1980s and 1990s. To the extent that future oil shocks are much less likely to disrupt consumer purchasing patterns, the deviations-from-trend equation appears to be the best forecasting equation.

This example shows that, while one could reasonably theorize that consumer spending patterns are based on long lags, that does not turn out to be the case empirically. That finding by itself does not invalidate the theory that long-run average or expected income is more important than current income as a determinant of consumption. Instead, it highlights several other results. First, the timing of many consumer purchases is dictated by the availability of credit. Second, expected income is more closely related to variables such as current stock prices and relative home prices than it is to lagged changes in real disposable income. As a result, long distributed lags do not work well in this equation. We will reexamine the forecasting efficacy of various consumption functions, with and without attitudinal variables, later in this text.

This case study also suggests that, while polynomial distributed lags are often useful, if they don't work, don't force them into the equation. In the case of functions for capital spending and the Federal funds rate (also bond yields), long lags are indeed important determinants. In the case of consumer spending, though, the lags are much shorter and, at least for discretionary purchases, the use of PDLs is inadvisable for building accurate forecasting equations.

Case Study 6: Capital Spending

Standard microeconomic theory states that, in equilibrium, the marginal product of labor is equal to the wage rate, and the marginal product of capital is equal to the cost of capital. However, whereas the wage rate is unambiguously defined in terms of dollars per hour, the cost of capital has a time dimension, since capital goods generally last for many years. In addition to the price of the capital good, one must also take into consideration the rate of interest, the cost of equity capital (stock prices), the rate of depreciation, and tax laws designed to affect capital spending – the accounting rate of depreciation and the rate of investment tax credit – as well as the corporate income tax rate. Also, the value of the marginal product of capital depends on the price of the product relative to the price of the capital good.

If the firm is operating on the constant part of its cost curve, the marginal product of capital is proportional to the average product of capital, which is output divided by capital stock. Under this assumption, then, optimal capital stock would be positively related to output and negatively related to the various components of the cost of capital.

This theory must be modified for several reasons. First, unlike labor, capital is "lumpy"; one cannot purchase half of a machine or a third of a building. Second, it takes time to fabricate a new machine or building. Third, firms may sometimes have excess capacity: thus even if output increases or the cost of capital declines, firms may not purchase new plant and equipment because the existing stock is adequate. Fourth, all the discussion so far has related to net investment; firms may replace existing plant and equipment as it wears out even if output has not increased.

These modifications make it difficult to estimate investment functions empirically. An additional complication, which is directly germane to the discussion in this chapter, is that the lag structures may differ for each of the variables. That is why PDLs are used to determine the lag structure. It is also possible that because of strong common trends, problems of multicollinearity will occur, which suggests using one of the methods of trend removal.

Differing lag structures may occur for each of the independent variables: GDP minus capital equipment, stock prices, the real Aaa bond rate, the effect of tax laws on investment, the price of capital equipment relative to the GDP deflator, and the relative price of oil. The latter term is included separately because the energy industry is more capital intensive than the rest of the economy, so when relative oil prices rise, the ratio of capital spending to GDP increases. That happened in both the late 1970s and the first half of the 1980s; when oil prices then fell, the ratio of capital spending to GDP dropped for several years.

At first one might expect the lag structure for all the terms to be about the same, since firms take both output and the cost of capital into consideration when determining their capital spending plans. However, that is not the case. The lags are much shorter for output than the cost of capital. There are several reasons for this that are beyond the scope of this discussion, but the overall argument can be summarized as follows. Consider two capital goods: one has a useful life of three years (personal computers, motor vehicles) and the other has a useful life of 20 years (electrical generating equipment, jet aircraft). The longer the economic life of the good, the more important the cost of capital becomes. A similar analogy can be drawn for consumers: the interest rate is much more important when buying a house than when buying a computer. Since a computer has a relatively short life, its purchase decision will be based primarily on the recent level of output. Thus the function for capital equipment can be viewed as a bifurcated model containing both the demand for both short-lived equipment, where output with a short lag is important, and long-lived equipment, where the cost of capital with a long lag is important.

That is indeed what we find, but one more adjustment must be made. During its lifetime, the investment tax credit was often used as a short-term policy variable to stimulate or reduce purchases of capital equipment: it was introduced in 1962, raised in 1964, suspended in 1966, reinstated in 1967, suspended in 1969, reinstated in 1971, raised in 1975, expanded in 1981, and modified in 1982 before being terminated in 1986. Depreciation allowances were also changed almost as frequently. That component of the user cost term has a shorter lag because it tended to affect the timing rather than the magnitude of capital spending.

Finally, note that the stock market variable serves a dual purpose. It is important with a short lag because it serves as a proxy for expected output. It is also important with a longer lag because it measures the equity cost of capital. As a result, the weights are quite high, then decline almost to zero, then rise again.

The levels function is given below. Including all of the individual terms for each of the PDLs would creates a listing of 63 separate terms, as given in the EViews program. That can be confusing for those not already familiar with equations containing PDLs. Hence the output is presented in summary form. The equation below presents the summary statistics for each of the independent variables; the sum of the coefficients and its standard error is given for each term estimated with PDLs. Figure 4.2 illustrates the lag structure for each of those variables.

$$
\begin{aligned}
PDE &= \underset{(13.8)}{505.6 - 291.0^{*}\,RPCG(-1)} - \underset{(10.6)}{0.046^{*}\,KSTOCK(-4)} + \underset{(12.4)}{0.128^{*}\,GDPXINV} \\
&\quad + \underset{(29.2)}{0.556^{*}\,SP500} - \underset{(1.7)}{102.8^{*}\,RCCE(-3)} - \underset{(19.8)}{27.1^{*}\,RAAA(-3)} + \underset{(8.9)}{69.4^{*}\,RPOIL(-8)} \\
&\qquad RSQ = 0.9984;\ SE = 6.65;\ DW = 0.92.
\end{aligned}
$$

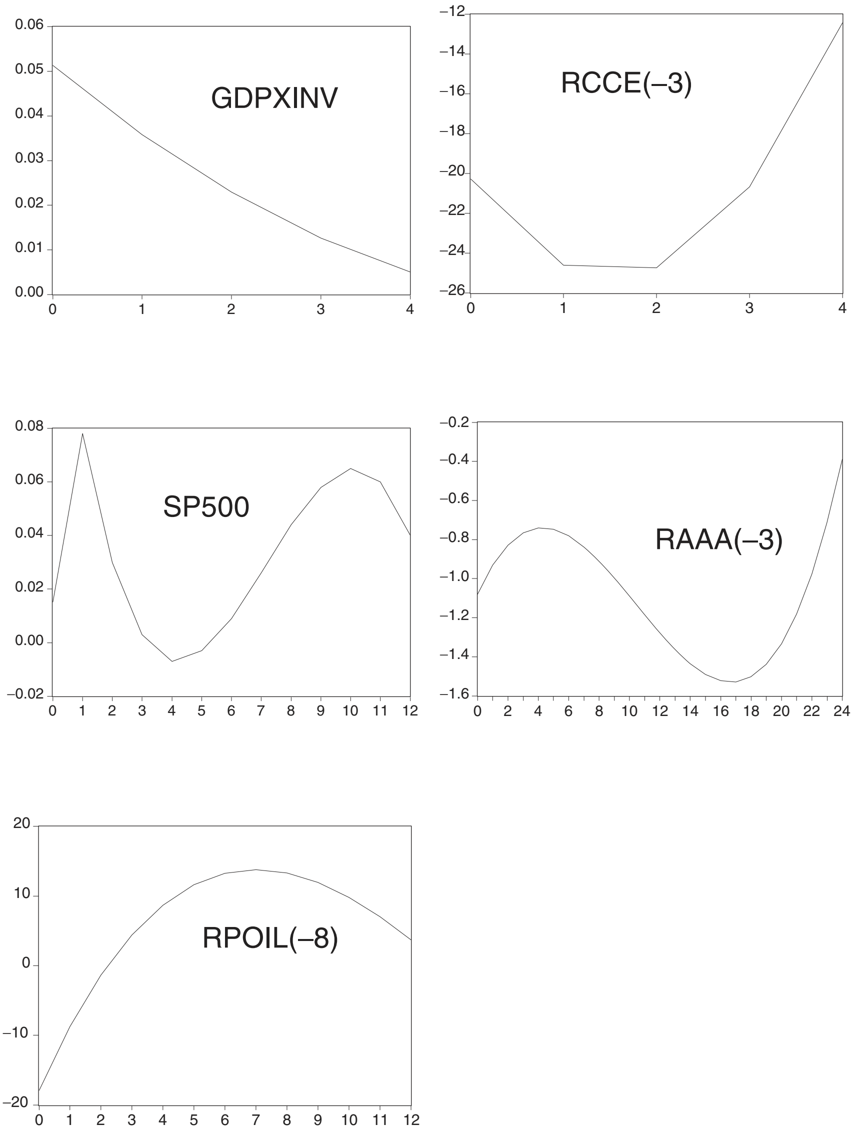

Figure 4.2 The lag structure for each of five variables with PDLs in case study 6.

In this equation, the dependent variable *PDE* is constant-dollar purchases of capital equipment. *RPCG* is the relative price of capital goods, *KSTOCK* is the capital stock of equipment, *GDPXINV* is real GDP excluding capital equipment, *SP500* is the S&P 500 index of stock prices, *RCCE* is the rental cost of capital for equipment, *RAAA* is the real Aaa corporate bond rate, and *RPOIL* is the relative price of oil.

This may appear to be a bewildering variety of lag structures. Some decrease monotonically; some start negative and turn positive; some decline and then recover; and some have two peaks. How does the researcher decide which lag structure is optimal?

The use of PDLs is more an art than a science. However, I offer the following general guidelines for determining the lag structure.

First, make an informed guess about the approximate length of the lag structure; if the guess turns out to be inaccurate, it probably won't affect the final result, but it will take more time. Since we know that capital spending decisions are based on information from several years, a reasonable starting point might be 3–4 years (12–16 quarters).

The default option for PDLs is usually a cubic polynomial constrained at the far end, which means that beyond the maximum length of lag specified, the coefficients are assumed to be zero. The value of the coefficients can also be constrained to zero at the near end of the lag structure, but that option is not generally used as an initial estimate. In EViews, the code for variable *R* (say) with this lag structure would be written as PDL(R, 12,3,2). *R* might be a level or percentage change, and it might also start with a lag. To estimate the percentage change of *R* starting with a three-quarter lag, the EViews code would be PDL(@PCH(R(–3)),12,3,2).

Usually, the initial result will not be satisfactory. Some of the terms in the lag distribution will be insignificant, and some will have the wrong signs. In some cases, the cubic term may not be significant, so a quadratic equation would be sufficient (saving one degree of freedom).

In general, if the individual terms in the PDL equation have a *t*-ratio of less than 2, it is best to shorten the lag structure until all remaining terms satisfy this criterion. If the terms at the end of the lag distribution appear to be highly significant, it is reasonable to lengthen the lag structure as long as the terms remain significant.

Sometimes the terms will be close to zero at the beginning of the distribution. That suggests starting that variable with a longer lag. Occasionally it also suggests using the constraint of a zero value for the coefficients at the near end of the distribution.

Those are the "simple" cases. Sometimes, the variable will start out significant, then drop into insignificance or have the wrong sign, and then become significant again. In other cases, the variable will change sign, with significant values that are both positive and negative. Should these results be kept or discarded?

Sometimes results of this sort are statistical flukes and should be discarded. Other times, however, they make economic sense. I have included one example of each type in the investment function. The coefficients of the stock market variable are highly significant, then drop off, then rise again because of the dual function of this variable. When the coefficients do turn negative, the *t*-ratios of those negative terms are less than 1.0, meaning they are not at all significant, so the structure is kept intact.

The other unusual case is the value of the coefficients of the relative price of oil, which switch from negative to positive. Here the lag distribution can be justified by economic theory. When oil prices rise (say) the initial impact is to reduce real growth and profitability for most firms, hence cutting investment. Eventually, though, higher oil prices stimulate investment in the energy industry, so the positive impact outweighs the negative effect.

The use of PDLs is not recommended unless the lag structure is thought to be fairly long and complex; for example, if the weights were thought to decline monotonically, that assumption could be entered into the regression without using a PDL. In the case of the investment function, it is difficult to estimate an equation that generates accurate predictions without using PDLs.

A few other comments on this equation seem appropriate at this point. Specifically, in spite of the apparently high R^2, which is fairly typical for time series with strong trends, we note that the *RCCE(–3)* term has a *t*-ratio of only 1.66; also, the *DW* is an uncomfortably low 0.92. Both these points are addressed below.

It would be tedious to repeat all this information for all six forms of the equation enumerated above; the interested reader may estimate these equations separately. The results are briefly summarized. In the logarithm equation, the *RCCE* term is much more significant (*t*-ratio of 6.1) but the *SP500* term drops out completely. The percentage change equation is dominated by random fluctuations and hence is of relatively little use; the capital stock, *RCCE*, *RAAA*, and *RPOIL* terms all drop out. Using the ratio of capital spending to GDP as the dependent variable restores the significance of all the variables, but the *SP500* term is now only significant with a short lag. All the terms are significant in the deviations-from-trend equation, and the *RPOIL* term almost drops out. Finally, switching to weighted least squares hardly changes the results relative to the levels equation.

Except for the percentage change equation, where almost all the terms drop out, the Durbin–Watson statistic indicates substantial autocorrelation. When the AR(1) adjustment is used, most of the terms are little changed except for the relative price of oil, which becomes insignificant. This raises the question of whether it should be dropped from the final form of the equation.

It is not immediately obvious from these comments which one of the five forms of the equation will generate the best forecast (since WLS hardly makes any difference, it is not considered further). The percentage change form can be ruled out because such a large proportion of the variance appears to be random fluctuations. The logarithm equation has too large a coefficient for capital stock and too small a coefficient for stock prices relative to a-priori expectations. The user cost of capital term is not significant in the levels equation, which does not seem to be a reasonable result. Hence the choice is between the ratio and the deviations-from-trend. Further experimentation reveals that when the ratio equation is "tidied up" to remove PDL coefficients that are insignificant or have the wrong lag, several other terms become insignificant, so the final choice is the deviation-from-trend equation. Since *DW* is quite low

CHAPTER 5

FORECASTING WITH A SINGLE-EQUATION REGRESSION MODEL

INTRODUCTION

Checking the structural stability of individual equations is one of the key steps in model building. Unfortunately, it is sometimes ignored, because the results are not so easy to fix. Autocorrelation, heteroscedasticity, or multicollinearity can generally be reduced by straightforward and relatively simple adjustments. However, if tests show that the equation is unstable, that often means starting all over again with a different specification.

It is often difficult to find a reasonable equation that meets all the statistical tests, including the stability of coefficients. Of course, even accomplishing this goal provides no guarantee that the forecasts will be accurate. However, if the equation is unstable, misspecified, or omits relevant variables, that almost guarantees the forecasts will not be accurate. Thus satisfying the tests discussed in this chapter is a necessary condition for successful forecasting, even though it is not sufficient.

We first discuss further tests for the residuals of the equation, checking for normality, autocorrelation, and heteroscedasticity, then turn to some of the methods that allow the researcher to check for the stability of the equation. These tests are then applied to a quarterly equation for new motor vehicles. An equation for housing starts is used to illustrate various alternative methods for adjusting the constant term. Finally, an equation for the dollar/yen crossrate is used to illustrate some of the pitfalls of multi-period forecasting.

5.1 CHECKING FOR NORMALLY DISTRIBUTED RESIDUALS

As noted in previous topics, the various goodness-of-fit statistics are valid only if the error term is normally distributed. Thus unless tests indicate this condition holds, the R^2 and t-statistics are likely to overstate the actual goodness of fit.

One of the useful characteristics of the normal distribution is that it is completely described by the mean and variance. However, that is not true for other distributions. Hence if the residuals are not normally distributed, higher-order moments, particularly skewness and kurtosis, might be significantly different from the normal distribution. That is the basis of the Jarque–Bera test, previously mentioned when histograms were discussed. To review briefly, that test statistic is given as:

$$JB = \frac{(T-k)}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \tag{5.1}$$

where S is skewness, K is kurtosis, T is the number of observations, and k is the number of variables.[1]

The probability and significance levels of JB are included in EViews, but it is usually obvious whether or not the residuals are normally distributed by looking at the histogram. If more than one outlier are three or more standard deviations from the mean, those observations are presumably not drawn from the same population as the rest of the sample period data. You have to decide whether that unusual situation is a one-time event that will not recur, in which case it probably will not affect the forecast, or whether it should be treated by adding another variable, using dummy variables, omitting the outliers, or including some nonlinear transformation.

Even if the residuals are normally distributed, that is only a starting point. In particular, the residuals could still be serially correlated, as is often the case for economic time series. Since this is one of the most common results for regressions estimated with time-series data, further tests are often warranted.

5.1.1 Higher-order Tests for Autocorrelation

The model builder has presumably already checked the Durbin–Watson statistic and determined whether or not significant first-order serial correlation exists. If it does, the equation may be altered by adding further variables, reducing or removing the trend, or including the first-order autoregressive adjustment AR(1). Using the lagged dependent variable on the right-hand side of the

[1] Most of the tests discussed in this chapter are also discussed in the EViews manual, ch. 7. The test was originally presented in Jarque, C. M., and A. K. Bera, "Efficient Tests for Normality, Homoscedasticity, and Serial Independence of Regression Residuals," *Economic Letters*, 6 (1980), 255–9.

equation is not recommended if the equation is to be used for multi-period forecasting.

Having made these adjustments, the residuals are now examined again. The procedure here follows the tests in EViews; while not all tests are included in this program, most others are duplicative or overlapping and are not discussed here.

It is sometimes useful to test for higher-order autocorrelation, especially if one is using quarterly or monthly data (it is unlikely to arise for annual data). While the DW test is obviously a useful starting point, it (i) tests only for first-order autocorrelation, (ii) does not work if the lagged dependent variable is included in the equation, and (iii) requires a constant term. For this reason, residuals are often tested with the Ljung–Box Q-statistic,[2] which is

$$Q_{LB} = T^*(T+2)\sum_{j=1}^{p} r_j^2 \Big/ (T-j) \tag{5.2}$$

where r_j is the jth order autocorrelation and T is the number of observations.

The Ljung–Box Q-statistic is often used with correlogram analysis, which is also available in EViews and other programs. That analysis shows two different correlations. The first is the correlation of ε_t with ε_{t-k}, where you pick the maximum value of k. The other is the partial autocorrelation, which shows the partial correlation of each coefficient when ε_t is regressed on $\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \ldots, \varepsilon_{t-k}$.

Ordinarily one would expect the value of the coefficients estimated in this equation to follow a geometrically declining lag. However, this test will also alert users to seasonality in the residuals.

Note that EViews includes two tests involving the Q-statistic. The first is for the levels of the residuals; the second is for the squared values of the residuals. Most of the time, both tests will generate the same results. However, sometimes the levels test will not show any autocorrelation while the squared test does, in which case it is likely that heteroscedasticity is also present. That does not happen every time, but it is a hint to look further.

The Breusch–Godfrey[3] test is an alternative test for autocorrelation. This test involves estimating a regression where ε_t is regressed on $\varepsilon_{t-1}, \varepsilon_{t-2}, \varepsilon_{t-3}, \ldots, \varepsilon_{t-k}$. The F-statistic is then calculated for this regression in order to see whether it is significantly different from zero. This is a large-sample test; the small-sample properties of the distribution are not known.

---

2 Ljung, G., and G. E. P. Box, "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 66 (1979), 265–70.

3 Breusch, T. S., "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17 (1978), 334–55; and Godfrey, L. G., "Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables," *Econometrica*, 46 (1978), 1293–302.

Breusch–Godfrey does allow the user to test whether there is any autocorrelation up to the lag limit specified in the test. The results are generally thought to be more accurate in situations where one uses the lagged dependent variable on the right-hand side of the equation. Since that method often leads to error buildup in multi-period forecasting, the Breusch–Godfrey test does not provide much additional information about improving multi-period forecast accuracy. It does, however, indicate that researchers may be fooling themselves if they think putting the lagged dependent variable on the right-hand side of the equation really eliminates autocorrelation.

5.1.2 Tests for Heteroscedasticity

Tests for heteroscedasticity are considered next. The ARCH test is based strictly on the values of the residuals, while the White test is based on the entire equation and hence is more general.

As noted previously, heteroscedasticity usually arises from one of two causes. The first is that the average size of the residuals increases as the size of the dependent variable increases. The second is that a few outliers dominate the regression estimates.

The simplest test is to correlate the residuals with each other. The only decision is how many lags to include. If the researcher chooses three lags, for example, estimate the equation

$$\varepsilon_t^2 = \beta_0 + \beta_1 \varepsilon_{t-1}{}^2 + \beta_2 \varepsilon_{t-2}{}^2 + \beta_3 \varepsilon_{t-3}{}^2 + V_t \tag{5.3}$$

and test whether the F-statistic from this regression is significant. One can also test whether the individual t-statistics are significant. This is known as the ARCH LM (Autoregressive Conditional Heteroscedasticity, Lagrangian Multiplier) procedure and is due to R. F. Engle.[4]

Halbert White[5] has developed a test to determine whether heteroscedasticity is present in the residuals; it can also be used to determine whether the equation is misspecified. First, estimate the basic regression

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Z_t + \varepsilon_t \tag{5.4}$$

and then calculate an auxiliary regression

$$\varepsilon_t^2 = \beta_1 + \beta_2 X_t + \beta_3 Z_t + \beta_4 X_t^2 + \beta_5 Z_t^2 + \beta_6 X_t Z_t + V_t \tag{5.5}$$

[4] Engle, R. F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50 (1982), 987–1008.

[5] White, Halbert, "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity," *Econometrica*, 48 (1980), 817–38.

If β_4, β_5, or β_6 is significant, heteroscedasticity is present, or the terms may not be linear. If the number of variables becomes large, the cross-terms (e.g., β_6) can be suppressed.

The White test is a more general test than merely determining whether heteroscedasticity of the residuals is present. If none of the coefficients is significant, it also suggests – although it does not verify – that the linear specification is correct.

5.2 Testing for Equation Stability and Robustness

In some cases, the residuals of the estimated equations will be normally distributed, with no autocorrelation and no heteroscedasticity, yet the equations themselves will generate poor forecasts. In many cases, this occurs because there has been a shift in the parameters of the underlying structural equation. In the worst possible case, the structure remained constant during the entire sample period but then radically shifted just as the forecast period started. There is no cure for that disease, and when it does happen, the forecasts will be inaccurate. However, that is a fairly unusual circumstance (although see case study 21 on page 346). Usually, any shift in the structure can be detected during the sample period. Under those circumstances, a battery of standard tests can be used to uncover this shift, permitting the model builder to adjust the forecasts accordingly.

5.2.1 Chow Test for Equation Stability

The Chow test[6] was one of the first regression diagnostic tests to be developed, and is still one of the most important. The idea is quite straightforward. Divide the sample period into two (or more) sub-periods. Calculate the regression during all of these periods, and then calculate the regression separately for individual sub-periods. The Chow test then compares the sum of the squared residuals obtained by fitting a single equation over the entire sample period with the residuals obtained by estimating separate equations over each sub-sample period. If the residuals are significantly different, the coefficients have probably shifted, which increases the probability the equation is unstable and the co-efficients will shift again in the forecast period, generating poor forecasts.

The Chow test measures the *F*-statistic of the difference between the total and sub-period squared residuals divided by the sum of the sub-period squared

[6] Chow, Gregory C., "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 52 (1960), 211–22.

residuals, adjusted for the number of observations and parameters. For the simplest case with two sub-periods, the F-test is given as

$$F = \frac{(\varepsilon^2 - (\varepsilon_1^2 + \varepsilon_2^2))/k}{(\varepsilon_1^2 + \varepsilon_2^2)/(T - 2k)} \tag{5.6}$$

where ε^2 is the sum of the squared residuals over the entire sample period, ${\varepsilon_1}^2$ and ${\varepsilon_2}^2$ are the sums of the squared residuals over the first and second sub-periods respectively, k is the number of parameters in the equation, and T is the total number of observations.

Sometimes when the Chow test shows unstable coefficients, the shift is caused by a readily identifiable economic or institutional factor, some of which have already been mentioned. For example, before the dollar was free to float, it was not possible to calculate a price elasticity for imports or exports because the dollar didn't change. Hence the coefficient would be different after 1971. Before the first energy crisis, the price of energy didn't change very much. Before banking deregulation, there were ceilings on interest rates; afterwards, the equations for both money supply and credit, as well as the importance of interest rates in the equations for aggregate demand, were far different. When wage and price controls were imposed, the impact of changes in demand on wages and prices was different than during normal periods. There are many such examples that must be treated explicitly, some of which can be handled with dummy variables. In other cases, the sample period can be truncated if it is likely that these circumstances will never reoccur, and hence are not relevant for current forecasting.

5.2.2 Ramsey RESET Test to Detect Misspecification

No mechanical test will ever turn a bad equation into a good one, just as in the stock market no mechanical method will ever make anyone rich or successful. On the other hand, various mechanical methods in the stock market can keep traders out of a certain amount of trouble, such as not buying a stock when it has just turned down after a long runup, insider selling is intensifying, the Fed is tightening, and profits are declining. To a certain extent, these mechanical tests merely quantify some of the more obvious errors that should have shown up simply by eyeballing the results.

Nonetheless, the Ramsey RESET test[7] can sometimes uncover errors that are not obvious by visual inspection. Indeed, Ramsey and Alexander[8] showed that

---

[7] Ramsey, J. B., "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society, Series B*, 31 (1969), 350–71.

[8] Ramsey, J. B., and A. Alexander, "The Econometric Approach to Business-Cycle Analysis Reconsidered," *Journal of Macroeconomics*, 6 (1984), 347–56.

the RESET test could detect specification error in an equation that was known a priori to be misspecified, but nonetheless gave satisfactory values for all of the more traditional test criteria – goodness of fit, high t-ratios, correct coefficient signs, and test for first-order autocorrelation.

The RESET test is designed to check for the following types of errors:

- omitted variables
- nonlinear functional forms (i.e., variables should be logs, powers, reciprocals, etc.)
- simultaneous-equation bias
- incorrect use of lagged dependent variables.

The Ramsey is developed as follows. In the standard linear model

$$Y_t = \beta_1 + \beta_2 X_{1t} + \beta_3 X_{2t} + \ldots + \beta_k X_{kt} + \varepsilon_t \tag{5.7}$$

consider the vector $\hat{\mathbf{Y}}_t$, which consists of the values fitted by the above equation. Ramsey now proposes the creation of a vector $\mathbf{Z}$, defined as

$$\left[\hat{\mathbf{Y}}_t^2 \; \hat{\mathbf{Y}}_t^3 \; \hat{\mathbf{Y}}_t^4 \ldots \hat{\mathbf{Y}}_t^k \right] \tag{5.8}$$

where the value of k is chosen by the researcher, and suggests that the powers of $\hat{\mathbf{Y}}$ be included in the equation in addition to all the other X_i terms that are already in the regression. The idea is that the various powers of the fitted values will reveal whether misspecification exists in the original equation by determining whether the powers of the fitted values are significantly different from zero.

The principal caveats can be grouped into three general categories.

(a) The Ramsey test is most likely to warn you if some of the independent variables should be included to powers greater than unity. However, that sort of misspecification does not occur very often. More often, the misspecification is due to the type of nonlinearity that occurs when one of the independent variables is a reciprocal, or the dependent variable increases at a faster rate during certain phases of the business cycle or in response to changes in economic conditions (such as a cost curve). Often, the Ramsey test will not discover such errors.

(b) Even if the Ramsey test signals that some variable(s) are omitted, it obviously doesn't tell you which ones.

(c) Problems associated with the lagged dependent variable are better fixed simply by deciding not to include such variables in the first place. In that sense, the Ramsey test will tell you something that should have already been known. In the case of the motor vehicle equation, which is described below, the RESET test shows that something is missing. It turns out to be fourth-order autocorrelation, which should have been detected in the previous tests.

It is always a good idea to run the Ramsey RESET test as an elementary diagnostic, but you cannot realistically expect it to find the "missing" variables for you. That is the job of the economist and forecaster.

5.2.3 Recursive Least Squares – Testing Outside the Sample Period

Econometricians realize that time-series regressions should be tested by omitting some of the data points. The Chow test provides just such a test. However, since the choice of break point chosen by the researcher is somewhat arbitrary, it would be better to examine the equation over time as sample points are added one by one. In the past, that used to represent a tremendous amount of regression time on the computer, and was seldom done. However, EViews provides the algorithms to combine these results and shows them on a single graph for each test, thus making the comparison and analysis much easier to absorb and analyze.

The EViews program performs six separate tests in this area; while some are more important than others, taken together they provide a comprehensive analysis of how the parameters change over time and whether the equation can reasonably be expected to remain stable over a reasonably long forecast period. All of these tests are illustrated in case study 7, the demand for motor vehicles.

Recursive Residuals

The program calculates values of the residuals in the period immediately after the end of the truncated sample period. For example, the equation would be estimated through 1975.4 and those coefficients would then be used to predict 1976.1. If that estimate falls outside two standard errors (as also calculated by the program), the test suggests the coefficients are unstable. This process is then repeated until the end of the full sample period.

CUSUM Test

The acronym stands for CUmulative SUM of the residuals. In this test, the cumulative sum of the residuals are plotted; hence the 2σ bands widen over time. If the equation generates one-period errors that are random, CUSUM will show the cumulative residuals remaining within their error bands. On the other hand, if the errors are cumulative (as would likely be the case if one were to use the lagged dependent variable on the right-hand side of the equation), they would increasingly lie outside the 2σ bands as time progresses.

CUSUM of Squares Test

If w_t is the recursive residual, then this test examines the ratio

$$s_t = \frac{\sum_{\tau=k+1}^{t} w_t^2}{\sum_{\tau=k+1}^{T} w_t^2}. \tag{5.9}$$

The only difference between the numerator and the denominator is that the numerator is summed only to an intermediate point t, whereas the denominator is summed to the final point T. Thus at the end of the sample period, s_t must equal 1. The question is then how s_t performs throughout the sample period. It should rise linearly from 0 to 1; if it falls outside the bands, that provides further evidence that the coefficients are unstable.

One-step Forecast Test

This is similar to the one-period-ahead residuals, except the graph also supplies probabilities to indicate where the biggest errors occur.

N-step Forecast Test

This test provides a series of Chow tests without having to specify any particular break point. The program calculates all feasible cases, starting with the smallest sample size consistent with estimating the equation.

Recursive Coefficient Estimates

This final test provides graphs of all the coefficients as the sample size increases from its minimum to the last observation. It enables one to determine whether the coefficients remain stable as more observations are added. In general, the coefficients tend to become more stable as the number of observations increases.

5.2.4 Additional Comments on Multicollinearity

As mentioned previously, there is no standard, mechanical test for multicollinearity. One has to rely on common sense when evaluating the parameter estimates.

It was previously noted that nonsensical parameter estimates and large standard errors were tipoffs that extreme multicollinearity is present. In addition, severe multicollinearity often is present if the coefficients change significantly as more data points are added. That is another reason why it is a good idea to estimate the equation over part of the sample period and then see whether the parameter estimates change significantly when additional observations are included.

Another test can be performed as follows. Suppose X and Y were highly collinear and also of the same size (as indicated by the correlation matrix), and suppose the coefficients were $0.4X$ and $0.75Y$. One possibility is to form a combination variable $Z_1 = 0.4X + 0.75Y$ and substitute that into the equation. Naturally that will generate the same result, except the goodness-of-fit statistics will be slightly better because one degree of freedom is added. Now vary the equation slightly by choosing $Z_2 = 0.3X + 0.85Y$. The other coefficients and the goodness-of-fit statistics should be just about the same. If they are not, the equation will probably be unstable in the forecast period and should be reformulated. Perhaps X or Y should be dropped, or the entire equation should be reformulated in percentage change instead of levels form.

If two of the independent variables are found to have an extremely high correlation with each other, ask yourself why. Are they really just two representations of the same economic phenomenon? If so, it is not necessary to include both terms in the equation. Or are they simply manifestations of the same common trend? I have dealt with that concern in previous chapters. Even if the equation remains in levels form, it is generally a good idea to check the stability of coefficients by calculating the corresponding percentage change function and see which variables remain significant.

Case Study 7: Demand for Motor Vehicles

We now turn to the regression for new motor vehicle (car and light truck) sales. The data are quarterly from 1959.1 through 1998.4. The key variables are real disposable income less transfer payments, the percentage change in real consumer credit, the relative price of gasoline, a dummy variable for auto strikes, stock prices, consumer credit, the unemployment rate, the yield spread between long and short-term rates, the ratio of two demographic variables – younger people between the ages of 16 and 24 buy more cars, and those between the ages of 45 and 64 buy fewer cars, ceteris paribus – and a dummy variable for the third quarter in 1985–7, described below. The complete equation is given in figure 5.1, and the actual and simulated fitted values are shown in figure 5.2.

All of the terms are highly significant. There is no autocorrelation in this equation, so no AR(1) adjustment is necessary. The reader should also rerun this regression without the DUMQ3 term to verify that all the terms remain significant and the equation is little changed except for a lower adjusted R^2 and higher standard error.

The first test is to whether the residuals are distributed normally. If the *DUMQ3* variable is included, that is indeed the case. On the other hand, if that variable is omitted, the probability that the residuals are normally distributed is zero to six decimal places. Clearly those observations are drawn from a different population, which justifies the use of the dummy variable.

Dependent Variable: MOTVEH
Method: Least Squares

Sample(adjusted): 1959:1 1998:4
Included observations: 160 after adjusting endpoints

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	–3.03	0.56	–5.4	0.000
INCOME EXCL TRANSFERS	0.00172	0.00015	11.4	0.000
@PCH(REALCRED,4)	8.65	1.11	7.8	0.000
CPIGAS(-2)/CPI(-2)	–2.81	0.56	–5.0	0.000
DSTR	0.88	0.21	4.3	0.000
UNEMPL RATE	–0.397	0.048	–8.3	0.000
@MOVAV(YLDSPRD (-2),4)	0.347	0.045	7.7	0.000
POP RATIO	19.1	0.92	20.7	0.000
SP500(-1)	0.00260	0.00059	4.4	0.000
DUMQ3	2.78	0.39	7.2	0.000

R-squared	0.959	Mean dependent var	12.23
Adjusted *R*-squared	0.958	S.D. dependent var	2.71
S.E. of regression	0.56	Akaike info criterion	1.75
Sum squared resid	47.4	Schwarz criterion	1.94
Log likelihood	–130	F-statistic	392
Durbin-Watson stat	1.76	Prob(*F*-statistic)	0.000

Figure 5.1 Tabular EViews output for the equation for motor vehicles in case study 7.

These outliers occurred in 1985.3 and 1986.3, and to a lesser extent in 1987.3. Car manufacturers, finding they had excess vehicles left at the end of the model year, used unprecedented discounts, rebates, and inexpensive financing to clear showroom floors. In retrospect, the auto industry miscalculated the demand for domestic vehicles because the dollar was so strong that many buyers switched to foreign models. However, using the value of the dollar won't work because the residuals are negative during most of the year but strongly positive during the third quarter.

If the dummy variable for these three years is included, the residuals are normally distributed. The coefficients do not change very much when this dummy variable is added, so the incentive contests in those years are unrelated to the other variables in the equation. One could argue that these incentive programs were not known in advance, so the forecast would not be improved by including such a variable. Yet it could also be claimed that with production running

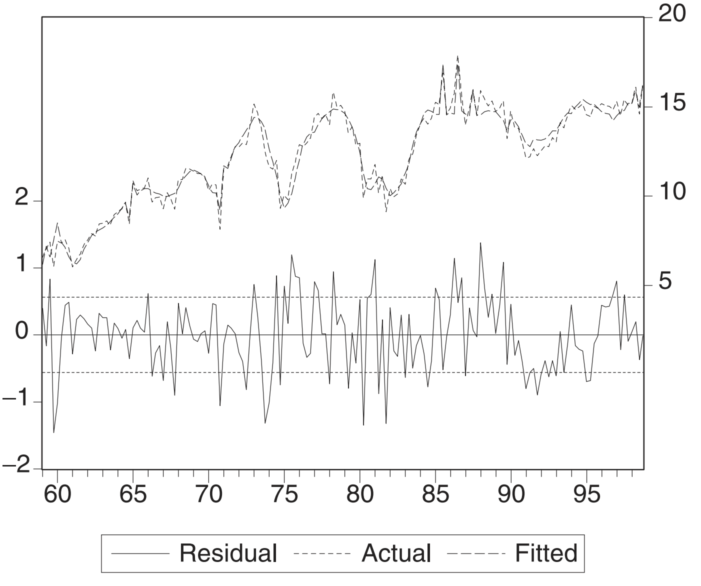

Figure 5.2 Graphical EViews output for the equation for motor vehicles.

ahead of sales for three years in a row, most ex ante forecasts would have adjusted for this discrepancy based on the grounds that the industry has never yet taken new cars back to the factory and melted them down.

The tests for autocorrelation and heteroscedasticity are inconclusive. The correlogram tests do not indicate positive autocorrelation. The Breusch–Godfrey test shows no significant autocorrelation with lags of 1 and 4, but some significant autocorrelation with lags of 2 and 3. The ARCH LM test shows no autocorrelation. The White test without cross-terms shows heteroscedasticity, but the test with cross-terms does not. That is fairly unusual; in most cases the two tests show the same results. Based on these results, autocorrelation and heteroscedasticity are probably not significant in this equation.

The Ramsey RESET test with lags of 2, 3, and 4 shows no misspecification. However, the reader should verify that if the population ratio variable is broken into two terms – those aged 16–24 and those aged 45–64 – the goodness-of-fit statistics of the equation are improved but the Ramsey equation indicates something is clearly amiss with the equation. While the test does not provide definitive answers, the most likely culprit is excessive multicollinearity when both population variables are entered separately.

The Chow breakpoint test shows a clear shift in the structural coefficients. The recursive coefficient tests show that most of the discontinuity occurs in the early 1970s, during the period of wage and price controls, the first energy crisis, and the first credit crunch. If the equation is reestimated starting in 1975, there are some minor changes in the coefficients, but the statistical problems autocorrela-

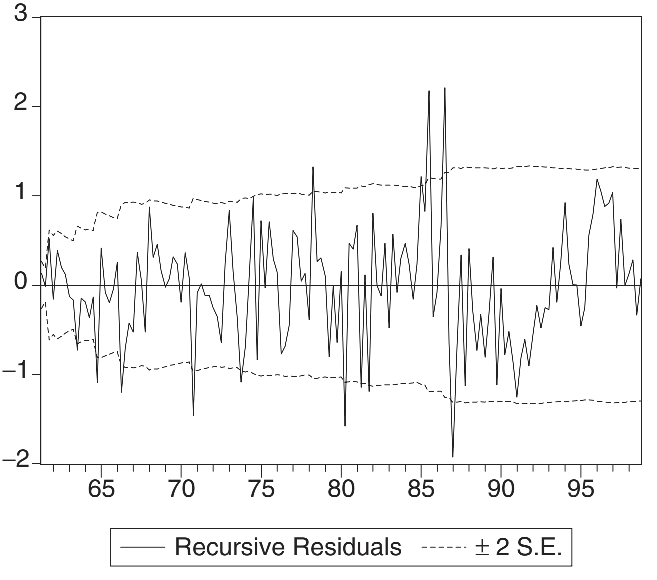

Figure 5.3 Recursive residuals test.

tion, heteroscedasticity, and shifting parameters disappear. Based on the 1975–98 sample period, the equation passes virtually all diagnostic tests.

Note that one cannot run the Chow test for most of the sample period because of the *DSTR* variable for auto strikes. Since there were no major strikes from 1970 through 1997, calculating the equation for during that period will result in the *DSTR* variable being entirely zeros, in which case the variance/covariance matrix won't invert. The way around this is to rerun the equation without the *DSTR* variable and then use the Chow test for various sample periods. However, the CUSUM tests provide a somewhat better way to handle this anomaly, so that method is used.

The first diagram of the recursive least squares testing shows what are called "recursive residuals," which are calculated as follows. The equation, which starts in 1959.1, is estimated through 1964.4. That equation is then used to predict 1965.1. The equation is then reestimated through 1965.1, and that version is used to predict the next quarter. The process continues through the end of the sample period. Note that in performing these tests, we must drop the *DUM3Q* variable because it is zero until 1985.3, If it were included, testing would be restricted to the period from 1986 to the present, and most of the information about shifting parameters would be lost.

The solid line in figure 5.3 shows the residuals calculated in this manner, and the dotted line shows the ranges defined by two standard errors. Points outside this range suggest the equation is unstable, particularly for the third quarters in the mid-1980s, although there is also some instability earlier. After 1987 there is no further instability.

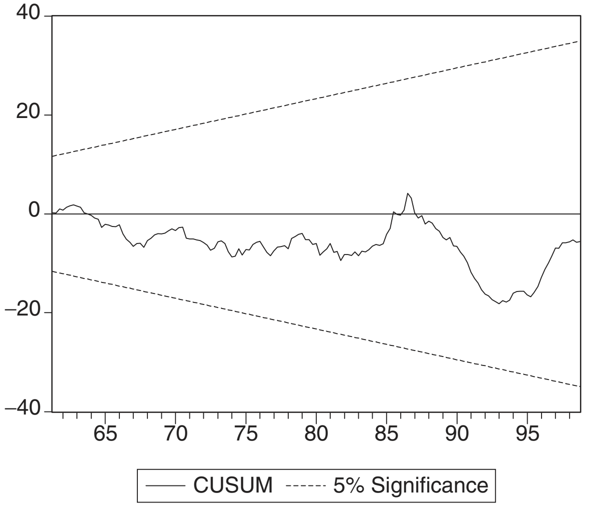

Figure 5.4 CUSUM of squares test.

Before drawing any firm conclusions, it is best to look at the next two tests, which are CUSUM and CUSUM of squares. CUSUM stands for cumulative sum of residuals, while CUSUM of squares is the cumulative sum of the square of the residuals. As the name implies, this is just the cumulation of the individual recursive residuals shown in the previous graph.

The recursive residuals test shows whether the forecasts are likely to be wide of the mark in any given quarter. But the CUSUM tests, illustrated in figures 5.4 and 5.5, show whether these one-time errors are just due to random factors or whether there is some systematic, long-lasting bias in the forecasts. For this equation, CUSUM shows that the errors do not accumulate over time, while CUSUM of squares shows structural instability before 1987.

The fourth test, known as the "one-step forecast test," is not shown separately, since it is more or less a repeat of the first graph. This test calculates the probability levels associated with each of these points. If the residual is well outside the two-standard deviation limit, the probability will be very close to 0 that the forecast point is drawn from the same population as the sample period. The fifth test known as the "*N*-step forecast test," also shows the same residuals as figure 5.3 and is not included separately. This test also shows that the probability of a stable sample is near zero until 1987. Before that date, the factors affecting car sales differed enough that the underlying equation changed almost every period. Some of these factors include the energy shocks, strikes, and deregulation of the banking sector, which led to alternative methods of credit allocation.

Finally, figure 5.6 shows the behavior of the individual coefficients except the constant term as sample points are added. Most of the coefficients are stable

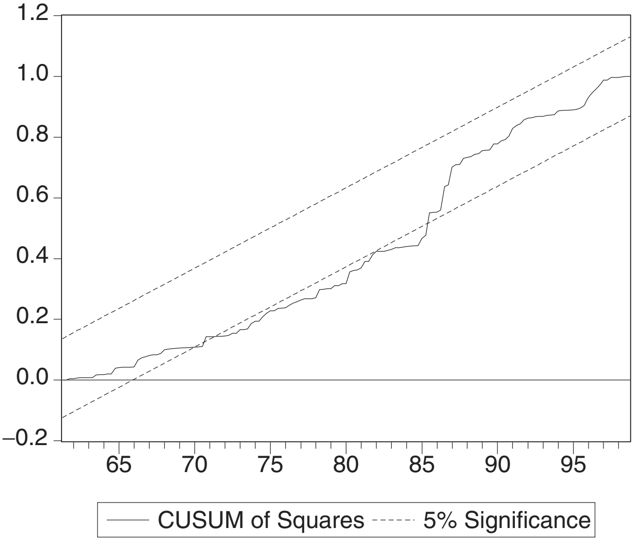

Figure 5.5 CUSUM of squares test.

after 1975; before then, some of them were buffeted by the first energy shock. Since then, the coefficients have remained generally stable except for the outliers in the third quarters of 1985, 1986, and 1987. The coefficients are shown in the order listed in figure 5.1.

This case study has illustrated the standard tests that should be used to determine whether the equation is statistically robust in the sample period and is likely to generate accurate forecasts. The main result we found was that the coefficients shifted during the early 1970s, and an equation estimated with post-1975 data provided much more robust statistics. However, that still does not answer the question of whether this equation would actually forecast well, since it requires accurate forecasts of the unemployment rate, stock prices, and consumer credit. Furthermore, in a complete macro model, it is also possible that the two-way causality in these variables would reduce forecasting accuracy.

5.3 Evaluating Forecast Accuracy

So far the standard error of the equation has been used to measure the accuracy of any given equation. When we move beyond the sample period, the corresponding measure is known as the root mean square error, or RMS forecast error,[9] defined as follows:

---

[9] While this is the standard measure of forecast errors, it sometimes gives ambiguous results. For a discussion of this, see Clements, Michael P., and David F. Hendry, *Forecasting Economic Time Series* (Cambridge University Press, Cambridge, UK), 1998, ch. 3, esp. p. 68.

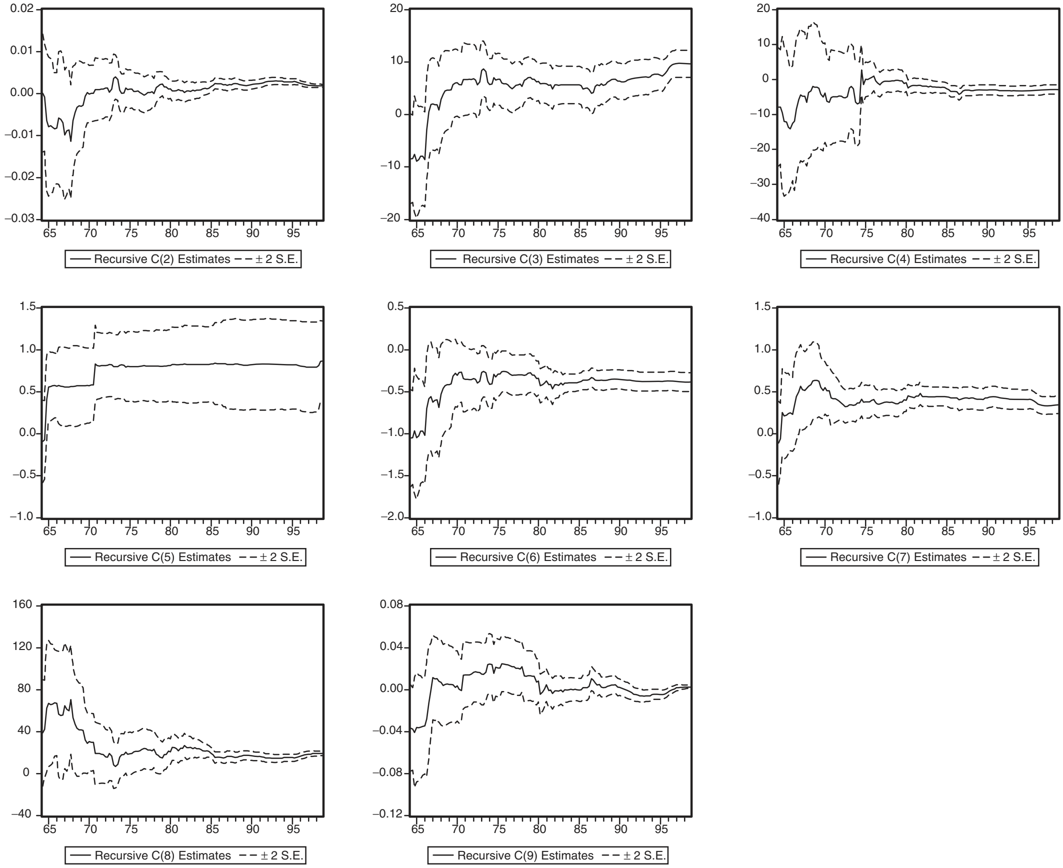

Figure 5.6 Recursive coefficient test.

$$RMSE = \sqrt{(1/T)\sum (Y_{ts} - Y_{ta})^2} \tag{5.10}$$

where Y_{ts} are forecast values of Y, and Y_{ta} are actual values.

A somewhat similar measure, which generally gives comparable results, is the "absolute average error" (AAE), which is simply the actual error without regard to sign. The RMS error takes into account the greater penalty associated with very large forecasting errors, and is a more statistically robust term. However, both terms are often used to evaluate forecast accuracy, and both are used in this text.

An equation that generates forecasts with a continuing bias may still be very useful if one can adjust for that bias. Suppose, for example, that all recent residuals have been about 5% above the actual numbers; that 5% factor can be incorporated in the forecast. That raises the question of why the predicted values are too high; perhaps the parameters should be reestimated. However, there may have been a structural shift in recent periods that is noticeable in the residuals but has not occurred for a long enough time to warrant the inclusion of an additional variable.

This point is now illustrated with a quarterly equation for housing starts. It will be shown that (i) while the equation adequately tracks all the major cycles in housing starts, the residuals are highly autocorrelated; (ii) the residuals in recent years have all been positive; (iii) adjusting the forecasts by the average size of the recent residual materially improves the forecast: (iv) adding additional variables to try and explain the recent residuals worsens forecast accuracy, and (v) adjusting for autocorrelation using the AR(1) method also materially worsens forecast accuracy. While these results are based on only one equation, this author can state after many years of actual forecasting that results of this sort are the rule rather than the exception.

Using the RMS may not be appropriate in all circumstances. In particular, individuals or firms might face an asymmetric loss function: if the predicted values are below the actual values, the firms may lose some sales, whereas if the predicted values are higher than the actual values, the firm may go bankrupt. It is also possible that in certain applications, such as financial markets, predicting the direction in which the particular market will move is just as important as predicting the magnitude of that move.[10] In other cases, trend forecasts may be less important than accurate predictions of turning points.

The individual forecaster can, if desired, construct other asymmetrical loss functions. Yet even in these cases, smaller forecasting errors are much preferred to larger ones. Thus in a book on practical business forecasting, the emphasis should be placed on the source of these forecasting errors, which can be grouped into three main categories:

(a) errors that occur from the random nature of the forecasting process – in the case of normally distributed residuals, these are accurately measured by the standard error of the equation over the sample period
(b) errors that occur because the underlying data generation function has shifted – as a subset, this includes exogenous influences that have not previously occurred
(c) errors that occur because the actual values of the independent variables are not known at the time of forecast.

There is a massive literature discussing point (a), but very little on (b) and (c). Yet as a practical matter, those errors are likely to be larger than the errors indicated by the standard error of the equation during the sample period. One of the major tasks of forecasters is to reduce the errors from (b) and (c) by adjusting the equations outside the sample period, and taking extraneous information into account. Since this is often done on an individual basis and involves trial and error, the statistical results are not very robust and, in some cases,

---

[10] For a discussion of asymmetric loss functions, see Clements and Hendry, pp. 102–4.

cannot be precisely quantified. However, that does not make them any less important. The remainder of this chapter includes some examples of how these errors might be reduced; the subject is then discussed in greater detail in chapter 8, after the material on time-series models has been presented.

5.4 The Effect of Forecasting Errors in the Independent Variables

There are many equations which, if correctly estimated and thoroughly tested, will generate forecasts that are well within the estimates indicated by the SE. In those cases, the econometricians have presumably done their work well, and no further comment is needed.

However, as this author can attest, there are often times when an equation that appears to be robust by all standard statistical tests generates very poor forecasts. The practical question is what to do in such situations. It is, of course, possible that the equation has been misspecified, so the only reasonable solution is to start all over again. In many cases, however, forecast accuracy can be improved by using some of the following tools. First, some econometricians suggest recalculating the equation with an AR(1) transformation. Second, it is often advantageous to adjust the constant terms based on recent residuals. Third, forecast error is often increased by faulty predictions of the independent variables; in some cases, using consensus forecasts may help. The advantages and disadvantages of these methods, together with several examples, comprise the remainder of this chapter.

It should also be noted that forecast error should be evaluated not only in comparison to the sample period error, but relative to errors generated from so-called naive models, which assume that the level or percentage change in a given variable is the same as last period, or – in a somewhat more sophisticated version – that the variable is a function only of its own lagged values and a time trend. Hence the forecasting record of naive models is also considered in these examples.

Case study 1 on page 90 presented an equation for the annual percentage changes in constant-dollar retail sales at hardware and building materials stores for the period from 1967 through 1998. The data series itself has a standard error of 6.6%, so if one assumed that hardware sales would rise at the average amount every year, the average forecast error would be 6.6%. A regression with current levels of the change in disposable income, housing starts, and the unemployment rate, and lagged changes in the Aaa corporate bond rate, explains 89% of the variance and reduces the standard error to 2.2%. It would appear this equation does indeed predict most of the change in the hardware sales.

Yet equations of this sort are always subject to multiple sources of possible forecast error. As noted above, the first is the random generation process that is reflected in the SE of the equation itself. The second test is the possibility that the structure will shift outside the sample period. The third is that the

Table 5.1 Errors generated by the following methods in the hardware store sales equation.

	Actual change (%)	Naive model	Errors		
			Within equation	Outside equation	With consensus values
1994	10.0	3.0	1.3	1.6	2.0
1995	1.4	8.6	1.8	1.9	3.4
1996	5.9	4.5	0.8	0.6	1.4
1997	5.5	0.4	0.2	0.1	0.7
1998	11.4	5.9	3.7	3.9	7.0
RMS		5.3	2.0	2.1	3.7

structure may remain unchanged, but forecasts for the unlagged values of the independent variables may be inaccurate.

Two major tests can be applied. The first one is to estimate the equation through a truncated sample period, ending in (say) 1993 and then forecast ahead, using the equation outside the sample period but inserting actual values of the independent variables. The second test is to use the consensus forecasts made each year for the percentage change in income, actual change in housing starts, and level of unemployment. Since the Aaa bond rate is lagged, it is known with certainty when the forecasts are made and hence does not contribute to any error for one-year forecasts.

SE for the 1994–8 period calculated with an equation estimated through 1993 is only 2.1%, virtually the same as the *SE* of the fitted residuals if the equation is estimated through 1998. This indicates stability of the equation. Of course this is never a perfect test, as it could be claimed the equation fits so well because we know what happened in the 1994–8 period and adjusted the sample period equation accordingly. In this particular case, however, the author actually used such an equation to predict hardware sales and can therefore warrant that the structural form of the equation did not change over that period.

The next test substitutes consensus forecasts for the actual values of income, housing starts, and the unemployment rate, and then recalculates *SE* for these five years. As shown in table 5.1, using predicted values for the unlagged independent variables boosted *SE* from 2.1% to 3.7%. These forecast errors are still smaller than the 5.2% error which would be generated by assuming the percentage change this year is the same as last year, but they are substantially higher than the ex post simulation errors. The error from the naive model is generated by assuming the change this year is the same as the change last year.

As shown in table 5.1, the difference in the RMS between the values when the sample period includes 1994–8 and excludes it is virtually nil. On the other hand, using the consensus instead of actual values almost doubles the error. This is not an atypical result, and illustrates one major reason why forecast errors are invariably larger than indicated by the sample-period standard error.

Because hardware sales are closely tied to housing starts and income, the structural stability is quite high, and most of the error stems from the inability to predict those variables. Examining errors from the equation for housing starts, on the other hand, shows that most of the error reflects a shift in the function rather than the inability to predict the independent variables; that case is considered next.

Case Study 8: Housing Starts

The equation for housing starts is estimated from 1959 through 1997. Data for housing starts are not comparable before 1959, and the sample period is truncated to allow several quarters to evaluate forecasting accuracy.

In the short run, housing starts are a function of both demand and supply variables, where supply in this case represents monetary factors: the cost and availability of credit. Short-run demand factors are represented by stock prices and the unemployment rate. Long-run demand factors are represented by demographic trends and the vacancy rate.

The availability of credit plays a key role; it is measured by the difference between the long-term and short-term interest rates, generally known as the yield spread. When the yield spread widens, credit becomes more easily available and financial institutions are more likely to lend money to both builders and homeowners. The change in the real money supply over the past two years is also an important variable. Because of changes in banking regulations, changes in the money supply were more important before 1973, while changes in the stock market were more important starting in 1980. Because housing starts are trendless, the ratio of stock prices to GDP and percentage changes of stock prices are used instead of levels. Similarly, the real money supply term is entered in percentage changes, and the number of people aged 20–24 is divided by total population. The dummy variable *DUM80* is included because of the discontinuity caused by starting the SP/GDP series in 1980.1. The equation is

$$
\begin{aligned}
&HST \\
&= 1.36 + \underset{(9.3)}{14.42} * POP20/POP - \underset{(10.3)}{0.148} * VACRAT(-5),4 - \underset{(5.5)}{0.062} * UN(-2),4 \\
&\quad + \underset{(8.5)}{0.165} * \%RM2,8 + \underset{(6.5)}{0.035} * \%RM2,4 * DUM73 + \underset{(12.3)}{0.140} * YLDSPRD(-2),8 \\
&\quad + \underset{(2.0)}{3.87} * [SP/GDP](-1) * DUM80 + \underset{(3.0)}{0.012} * \%SP,4 - \underset{(2.8)}{0.271} * DUM80
\end{aligned}
$$

$$RSQ = 0.884;\ SE = 0.110;\ DW = 1.01$$

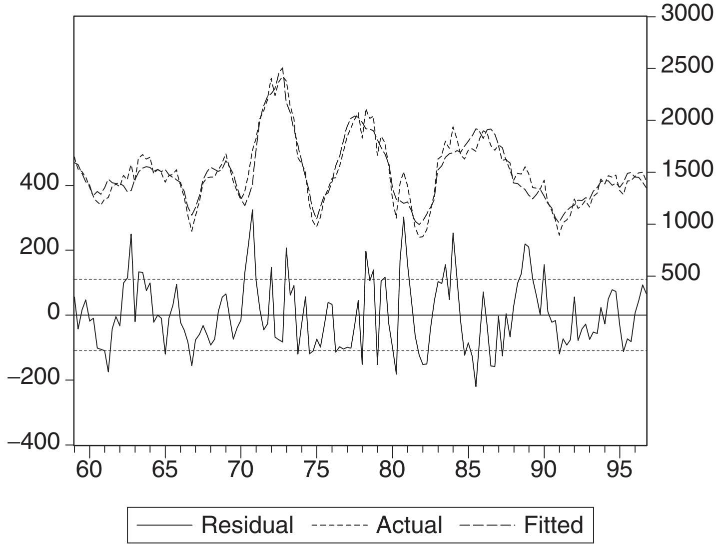

Figure 5.7 Residuals for the housing starts equation.

where *HST* is housing starts in millions, *POP20/POP* is the ratio of population aged 20–24 to total population, *VACRAT* is the vacancy rate for rental properties (the (–5),4 means it starts with a five-quarter lag and extends back an additional four quarters), *UN* is the unemployment rate, *RM2* is real money supply (M2 divided by CPI), *YLDSPRD* is the spread between the Aaa corporate bond rate and the Federal funds rate, *SP/GDP* is the ratio of S&P 500 stock price index to GDP, *DUM73* is a dummy variable (1 through 1972.4, 0 thereafter), and *DUM80* is a dummy variable (0 through 1979.4, 1 thereafter). The numbers after the commas indicate the length of lag. In levels terms, these lags represent the length of the moving average; in percentage terms, they indicate the length of lag for the percentage change.

The equation appears to track the data very well, as shown by the comparison of the actual and fitted values in figure 5.7. But how well does it forecast?

That all depends on what measuring stick is used. The first and easiest step is to calculate the difference between actual and estimated values for 1997.1 through 2000.4 generated by this equation. That period is outside the sample period, and appears to give very good results, as shown in table 5.2. However, that is not a true test for two reasons. First, this equation benefits from the hindsight of including the stock market variable. Second and more important, any such calculations assume that the values of all the independent variables – including the stock market – are known, which is obviously not the case. In particular, hardly anyone at the end of 1996 correctly predicted the actual increase in the stock market over the next three years.

For ease of exposition, the remaining forecasts are shown only for the annual totals. Table 5.3 shows the forecasts using the above equation when the actual

Table 5.2 Actual housing starts and forecast errors, with and without AR(1) transformation, using actual values for all independent variables.

Quarter	Actual	Error	Error with AR(1)
1997:1	1.459	0.098	0.062
1997:2	1.473	0.110	0.085
1997:3	1.457	0.075	0.046
1997:4	1.515	0.051	0.028
1998:1	1.585	0.079	0.034
1998:2	1.570	–0.003	–0.048
1998:3	1.637	0.012	–0.029
1998:4	1.701	0.077	0.009
1999:1	1.760	0.139	0.080
1999:2	1.591	–0.068	–0.097
1999:3	1.663	0.005	–0.016
1999:4	1.689	0.080	0.056
2000:1	1.732	0.137	0.117
2000:2	1.605	0.032	0.013
2000:3	1.527	–0.020	–0.034
2000:4	1.550	–0.004	–0.023

Table 5.3 Housing starts: comparison using actual values, consensus, and a naive model.

	Actual starts	Using eqn	AAE	With AR(1)	AAE	Consensus	AAE	Naive model	AAE
1997	1.48	1.39	0.09	1.42	0.06	1.40	0.08	1.35	0.13
1998	1.62	1.58	0.04	1.63	0.01	1.44	0.18	1.48	0.14
1999	1.67	1.64	0.03	1.67	0.00	1.53	0.14	1.62	0.05
2000	1.60	1.57	0.03	1.59	0.01	1.54	0.06	1.67	0.07
Average AAE			0.05		0.02		0.12		0.10

values of the independent variables are known, compared to the consensus forecast and a naive model. Table 5.4 shows the comparison when the consensus forecasts of the independent variables are used. The naive model assumes starts this year are the same as last year. In these tables AAE stands for absolute error.

The equation appears to do much better than the consensus forecast – but only if one knows the actual values of interest rates, the yield spread, growth in the money supply, and stock prices. Thus a more realistic comparison would use the consensus forecasts for these variables. Since consensus estimates for monetary variables are not available, it has been assumed that the yield spread

Table 5.4 Housing starts: comparison using consensus values in the equation.

	Actual	Using eqn	AAE	With AR(1)	AAE
1997	1.48	1.35	0.13	1.38	0.10
1998	1.62	1.40	0.22	1.43	0.19
1999	1.67	1.44	0.23	1.45	0.22
2000	1.60	1.45	0.15	1.46	0.14
Average AAE			0.18		0.16

Table 5.5 Housing starts: equation without stock market terms.

	No AR(1)	AAE	With AR(1)	AAE
1997	1.24	0.26	1.32	0.16
1998	1.36	0.26	1.43	0.19
1999	1.34	0.31	1.42	0.25
2000	1.27	0.33	1.32	0.28
Average AAE		0.29		0.22

remained at 1½%, real money supply rose 4% per year, and stock prices rose 12% per year. Entering these results in the above equation generates quite different results (tables 5.4 and 5.5).

Further analysis of these errors reveals that most of the mistake occurred because of the inability to predict the rapid rise in the stock market. In 1996, most economists did not include stock prices in the housing start equation; previously, the relationship had not been very robust, and indeed the significance levels of the two stock market terms are much smaller than the other terms; by 2000, these terms have become much more important.[11] Thus to generate a true ex ante forecast, it makes more sense to reestimate the equation through 1996 without the stock market terms and see how well it performs. The AAEs for these equations are so large as to render them useless for forecasting, since they are much bigger than the AAE for the naive model.

Note what has happened here. When actual values of the independent variables are used, the AAE is only 0.05 million starts. When the consensus values are used, that error rises to 0.18 million, and when the stock market term is omitted, it rises to 0.29 million. By comparison, the naive model AAE is only

[11] It remains to be seen whether this is a long-term stable relationship. When the stock market plunged in late 2000 and early 2001, housing starts improved because interest rates fell.

Table 5.6 Housing starts: using the equation without stock prices plus constant adjustment.

	Actual starts	Predicted	AAE
1997	1.48	1.38	0.10
1998	1.62	1.60	0.02
1999	1.68	1.61	0.07
2000	1.60	1.60	0.00
Average AAE			0.05

0.10 million. It might appear that the inability to predict the independent variables accurately vitiates the econometric approach. However, it is not yet time to give up.

It is possible to use the information in the residuals to adjust the forecasts each year. To do this, calculate the residuals – actual minus predicted values – each year for the previous four quarters; initially, the residual values are used for 1996. That average residual is then added to the 1997 forecasts generated using consensus estimates for the independent variables. The residuals for 1997 are then calculated, and added to the 1998 forecasts calculated from the estimates. The same procedure is used for 1999 and 2000. In other words, the forecasts for each year are adjusted by the errors of the previous year. When that relatively simple procedure is used, the results are much improved, as shown in table 5.6. The AAE is then seen to be much better than the consensus forecast – and as good as an equation using the stock price terms and the actual values for the independent variables.

The next two figures illustrate these various comparisons for quarterly values. In figure 5.8, actual housing starts are compared with the single-equation estimates including stock prices with and without AR(1) using the *actual* values of the independent variables, and using the consensus values of the independent variables. Clearly, most of the error stems from the inability to predict stock prices, money supply growth, and the yield spread.

Figure 5.9 shows the comparison of actual housing starts to forecasts made using an equation without the stock price terms and (i) no adjustment, (ii) the AR(1) adjustment, and (iii) the constant adjustment described above. There is no corresponding line for the equation with AR(1) and constant adjustments, because with AR(1) the residuals are random and hence any constant adjustment would be zero. It is clear that the improvement from using constant adjustments is much greater than the improvement from using the AR(1) transformation.

This case study is typical of the problems forecasters face when generating true ex ante predictions. In fact, the equation estimated in 1996 would have been misspecified by not using stock market terms – but that was not known

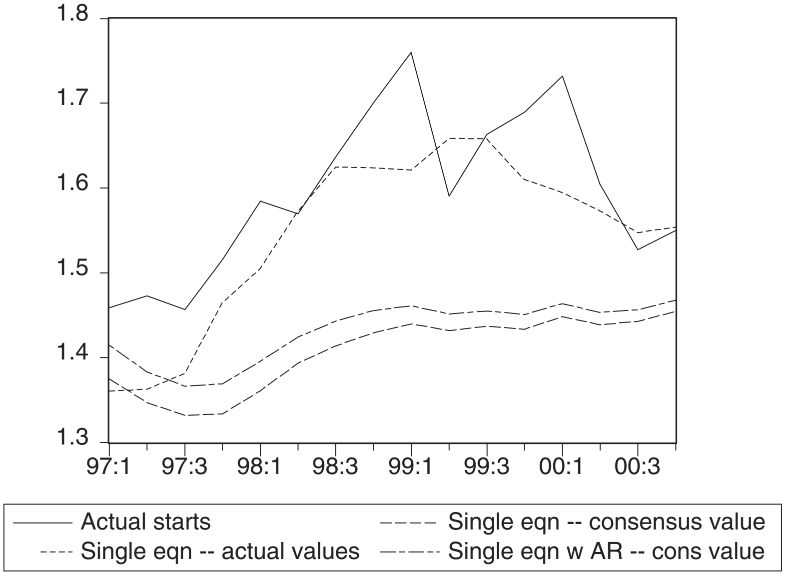

Figure 5.8 Actual and predicted levels of housing starts, comparing forecasts with actual and consensus estimates of the independent variables.

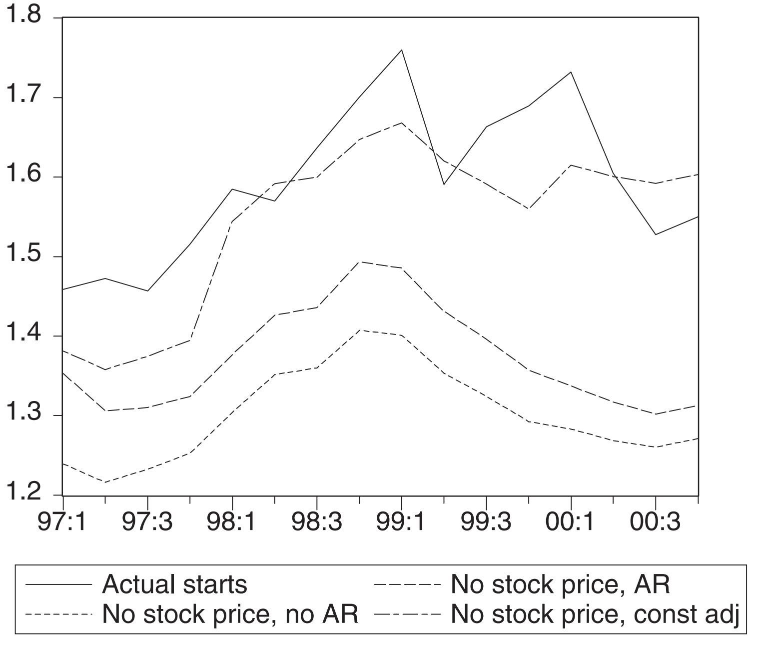

Figure 5.9 Actual and predicted levels of housing starts, using different assumptions about the independent variables and adjustment of the constant term.

at the time. The question is how to improve the forecasts in spite of not having perfect hindsight.

As these results show, using an AR(1) transformation helps very little; in other similar cases, it does not help at all. Using the consensus forecasts, while slightly better than a naive model, does not help very much either. The most successful method of improving forecast accuracy in this case, given a misspecified equation, is to adjust the constant term by the average of the previous year's residuals. That is a particularly helpful method when all the residuals have the same sign and approximately the same magnitude. Forecast accuracy can be significantly improved with this method, as will be shown further in Part IV.

5.5 Comparison with Naive Models

So far we have tested whether the residuals of the equations are normally distributed, whether autocorrelation or heteroscedasticity exists, and whether the coefficients are stable. However, even if all these conditions are met, the equation will not be very useful for forecasting if it does not produce smaller errors than a naive model that does not depend on any structural parameters. Two types of naive models are considered here. The first says the level – or change – this period is the same as last period. The second says that the variable is a function of its own lagged values.

5.5.1 Same Level or Percentage Change

If a variable has no trend – such as interest rates, capacity utilization, or the ratio of inventories to sales – the simplest naive model says the value of that variable this period is the same as last period. If the variable has a significant trend, that naive model says the percentage change this period is the same as the percentage change last period.

At first glance it might seem that any forecast at all could outperform a naive model; after all, that is the same as saying that R^2 in an equation is zero; none of the variance can be explained. Yet in some cases this is simply not the case. So far, no one has developed a model that will accurately predict the amount, or even the direction, that the stock market, interest rates, or foreign exchange rates will change the following day. A model that predicted as little as 10% of the variance would be extremely useful, but no such model has yet been found. Even if some formula were discovered that happened to work on some previous set of data, it would quickly become obsolete as traders and speculators rushed to take advantage of that information.

Nonetheless, it is easy enough to run so many regression equations that eventually some set of variables will turn up as highly significant. Such data mining exercises tell us nothing about the forecasting efficacy of such equations. Most

Table 5.7 Blue Chip Consensus forecasting record for percentage change in real GDP.

Forecast year	Consensus	Actual	Actual error	Naive error
1977	5.1	4.9	0.2	0.7
1978	4.3	5.0	−0.7	−0.1
1979	2.1	2.9	0.8	−2.1
1980	−2.0	−0.3	−1.7	−3.2
1981	0.7	2.5	−1.8	2.8
1982	0.3	−2.1	2.4	−4.6
1983	2.5	4.0	−1.5	6.1
1984	5.3	6.8	−1.5	2.8
1985	3.3	3.7	−0.4	−3.1
1986	3.0	3.0	0.0	−0.7
10-year absolute average error			1.1	2.6
1987	2.4	2.9	−0.5	−0.1
1988	2.2	3.8	−1.6	0.9
1989	2.7	3.4	−0.7	−0.4
1990	1.7	1.2	0.5	−2.2
1991	−0.1	−0.9	0.8	−2.1
1992	1.6	2.7	−1.1	3.6
1993	2.9	2.3	0.6	−0.4
1994	3.0	3.5	−0.5	1.2
1995	3.1	2.3	0.8	−1.2
1996	2.2	3.4	−1.2	1.1
1997	2.3	3.9	−1.6	0.5
1998	2.5	4.1	−1.6	0.2
12-year absolute average error			1.0	1.2
22-year absolute average error			1.0	1.8

of the time, however, it is difficult to perform actual ex post tests of forecasting accuracy.

Key macroeconomic variables represent one of the few cases where a documented track record of sufficient length exists, permitting the comparison of naive models with actual forecasts, as opposed to those generated from regression models when the results are already known.

The results for the Blue Chip Economic Indicators consensus forecasts are shown here (see tables 5.7 and 5.8), since they are the most widely used and

Table 5.8 Blue Chip Consensus forecasting record for change in CPI inflation rate.

Forecast year	Consensus	Actual	Error	Naive error
1980	11.0	13.5	–2.5	–2.2
1981	11.2	10.3	0.9	3.2
1982	7.8	6.2	1.6	4.1
1983	5.0	3.2	1.8	3.0
1984	5.0	4.3	0.7	–1.1
1985	4.2	3.6	0.6	0.7
1986	3.6	1.9	1.7	1.7
1987	3.2	3.6	–0.4	–1.7
8-year absolute average error			1.3	2.2
1988	4.2	4.1	0.1	–0.5
1989	4.7	4.8	–0.1	–0.7
1990	4.1	5.4	–1.3	–0.6
1991	4.8	4.2	0.6	1.2
1992	3.3	3.0	0.3	1.2
1993	3.1	3.0	0.1	0.0
1994	2.8	2.6	0.2	–0.4
1995	3.3	2.8	0.5	0.2
1996	2.8	3.0	–0.2	0.2
1997	2.9	2.3	–0.6	0.7
1998	2.2	1.6	–0.6	0.7
11-year absolute average error			0.4	0.6
19-year absolute average error			0.8	1.3

best documented of the consensus forecasts.[12] The actual forecasts, taken from tables prepared by Blue Chip, present estimates for both real growth and inflation, together with the naive models that say the rate of growth in real GDP and the CPI this year will be the same as last year. The consensus forecasts are those made at the beginning of January each year by 50 leading forecasters.

While these statistics are often calculated using the RMSE, they are shown here with the AAE, which is the statistic reported by Blue Chip; in any case,

[12] These are published and released on approximately the tenth day of each month by Panel Publishers, a division of Aspen Publishers, Inc., a Wolters Kluwer Company (Alexandria, VA).

the comparative results are not changed. The AAE for real growth is 1.0% for Blue Chip, compared with a naive model error of 1.8%. For inflation, the comparison also favors the Blue Chip consensus, with an AAE of 0.8% compared with a naive model estimate of 1.3%.

This test, however, is not conclusive. To understand the next point, it is necessary to understand the difference between quarterly average and annual average forecasts. This question has arisen in numerous presentations delivered by this author, and merits a brief explanation. The percentage change on an annual average basis, which are the numbers quoted in Blue Chip and many other sources, compares the average for the four quarters of this year with the four quarters last year. The percentage change on a quarterly average forecasts compares the fourth quarter of this year with the fourth quarter of last year.

The point can be expanded with an illustration. Suppose you are given data for the previous eight quarters and are asked to predict the following year based on the data in case A and the data in case B

Year/qtr	*Case A*	*Case B*
1.01	101	101
1.02	102	102
1.03	103	103
1.04	104	104
2.01	105	105
2.02	106	106
2.03	107	105
2.04	108	104

In case A, assume the forecaster knows nothing except that, over the long run, this variable grows at 4% per year. In that case, the forecast for the third year, using the first set of data, would probably be 110.5, showing (about) a 4% rate of growth; this simple example ignores the effects of compounding. In case B, the economy may be heading into a recession. Since recessions usually last 3–4 quarters, one reasonable forecast might be 103, 102, 103, 104, for an average of 103 for the year and a 2% decline from the previous year. However, suppose the forecaster is not astute enough to realize a recession has started and simply plugs in the 4% growth formula for the following year, giving figures of 105, 106, 107, and 108. That leads to an annual average of 106.5, which still shows only a 1½% growth rate for the year. It appears the forecaster has accurately predicted at least part of the slowdown, when in fact he or she merely took into account those data that were already known. By comparison, on a quarterly average basis, those data would not be known in advance. In plain English, predicting annual average changes provides forecasters with a "head start."

Yet this test is not conclusive either. The consensus forecast, like any other forecast, is based on the most recent data published by the government, but often these series are substantially revised in the years ahead. BEA, which publishes these figures, has calculated that the AAE for real GDP caused by revisions from the first "advance" estimate to the final figure is 1.4%, which is in fact larger than the AAE generated by the consensus forecast. About half of that differential is due to more complete data for the previous quarter, and about half is due to changes in methodology (such as including business purchases of software in capital spending). Thus even if the consensus forecasts were perfect at the time they were made, they might contain substantial errors when those forecasts are compared with revised data.

For these reasons it is actually not possible to construct a "clean" test that determines how well the consensus forecast has performed relative to a naive model or any other similar standard. Based on experience in generating macro forecasts since 1963, it is this author's opinion that the consensus forecast for real growth is significantly better than a naive model, the consensus forecast for inflation is slightly better, and the consensus forecast for interest rates is no better at all. In what could be considered a market test, a service offered by Blue Chip to provide consensus interest rate forecasts was withdrawn after a few years, whereas the basic Blue Chip forecast service remains popular with subscribers.

5.5.2 Naive Models Using Lagged Values of the Dependent Variables

The residuals in the housing starts equation discussed above have a high degree of autocorrelation. An alternative approach would be to estimate housing starts this quarter as a function of housing starts in previous quarters, eliminating any coefficients with *t*-ratios that are less than unity. Note this is a purely mechanical approach; no attempt is made here to estimate a structural equation.

While the most important variable is housing starts lagged one quarter, housing starts with lags of three and six quarters are also significant. The *SE* of 0.120 is slightly higher than the *SE* of 0.110 for the structural equation. The forecast errors are given in table 5.9; the AAE of 0.090 is higher than for either

Table 5.9 Absolute values of forecast errors from naive housing starts model.

1997.1	0.081	1998.1	0.007	1999.1	0.214	2000.1	0.060
1997.2	0.058	1998.2	0.017	1999.2	0.090	2000.2	0.069
1997.3	0.079	1998.3	0.078	1999.3	0.017	2000.3	0.140
1997.4	0.021	1998.4	0.173	1999.4	0.043	2000.4	0.090
1997	0.060						
1998	0.069						
1999	0.140						
2000	0.090						

4-yr average forecast error = 0.090

the equation with stock prices or the adjustment of the previous yearly residuals. At least for housing starts, the naive model does not work.

5.6 Buildup of Forecast Error Outside the Sample Period

We have seen that it is important to distinguish between unconditional forecasts, where all of the dependent variables are known with certainty – because they are non-stochastic or are lagged – and conditional forecasts, where the independent variables must also be predicted. Most practical business forecasts are of the latter variety, whereas the standard statistical tests are developed for the former case. This section provides further commentary on the additional error that is likely to occur in the forecast period even if the underlying structure of the equation remains unchanged. Three cases are discussed: error as the dependent variable moves further away from the mean; error because the values of the independent variables are not known; and error buildup in multi-period forecasting.

5.6.1 Increased Distance from the Mean Value

The algebra is fairly cumbersome without matrix notation for the multivariate case, so to illustrate this principal, consider the bivariate case, first for unconditional and then for conditional forecasts. Starting with the standard bivariate model[13]:

$$Y_t = a + bX_t + \varepsilon_t \tag{5.11}$$

one can generate a forecast

$$\hat{Y}_t = \hat{a} + \hat{b}X_{t+1}. \tag{5.12}$$

Remember that a and b are not known, but are estimated, so there are some errors attached to these estimates. Then

$$e_{t+1} = \hat{Y}_{t+1} - Y_{t+1} = (\hat{a} - a) + (\hat{b} - b)X_{t+1} - \varepsilon_{t+1}. \tag{5.13}$$

Note the two sources of error: one because of the random nature of the parameters that are being estimated, and the other because of the error term ε_{t+1}. After doing the arithmetic we have

[13] For further discussion of this point, see Pindyck, Robert S., and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th edn (Irwin McGraw-Hill, Boston), 1998, pp. 204–9 and 221–3.

$$s_f^2 = s^2\left[1 + 1/T + (X_{T+1} - \overline{X})^2 \Big/ \sum (X_t - \overline{X})^2\right] \tag{5.14}$$

where s_f^2 is the standard error in the forecast period and s^2 is the standard error in the sample period.

The second term in the square brackets, $1/T$, is a small-sample adjustment that disappears as the sample size increases. The third term is of more interest. In general, it means that the more the forecast value moves away from the sample period mean, the larger the forecast error. Hence the error is more likely to increase for a series with a strong trend, such as consumption or stock prices, than for a series with no trend, such as housing starts or interest rates.

To see how large that error might be, consider the consumption function. Over the 50-year period from 1949 through 1998, consumption had a standard deviation of \$1200 billion, which means the denominator of the above term is [50*(1200)^2], or \$72,000,000 billion. In 1999, the difference between the actual and mean value of consumption was about \$2600 billion, so its square would be \$6,760,000 billion, or about 9%. When $1/T = 2\%$ is added, the variance in the forecast period would be about 11% more than indicated by the sample period statistics – provided that the value of income is known with certainty (remember, this is just the bivariate case).

This source of error can be compared with a trendless series. The series for housing starts has a standard error of 0.3 million, so it has a variance of 0.09 million times 40 annual observations, or 3.6 million. As shown above, the error in forecasting housing starts is about 0.15 million, so the numerator is 0.02, or about ¼% of the denominator. Thus in the case of a variable without any trend, the increase in error during the forecast period from this source is minuscule.

5.6.2 Unknown Values of Independent Variables

We next consider the more realistic case of conditional forecasting. The same bivariate forecasting model is used, except that the value of X_t is not known at the time of forecast, so one also has to consider the error made in forecasting this variable. This time, add the assumption that $\hat{X}_{t+1} = X_{t+1} + u_{t+1}$, where u_{t+1} is the error in predicting X_{t+1}. Again skipping the algebra, the formula is

$$s_f^2 = s^2\left[1 + 1/T + \left\{(X_{T+1} - \overline{X})^2 + s_u^2\right\} \Big/ \sum (X_t - \overline{X})^2\right] + \beta^2 s_u^2. \tag{5.15}$$

In the case of the consumption function, the error in predicting income would generally be about the same as the error in predicting consumption when income was known. Also, β would be fairly close to unity. As a result, the term $\beta^2 s_u^2$ could turn out to be almost as large as s^2, in which case the forecast error would be almost doubled because the value of the independent variable was not known. Furthermore, if the number of variables in the equation is expanded, the error accumulation can become quite substantial. The result is not generally additive because there are usually some negative covariances – some offsetting errors – but nonetheless can be quite large relative to the SE.

We have already seen how the forecast error almost doubled in the equation for hardware store sales when the consensus values were substituted for the actual values of the unlagged independent variables. This formula explains how such an error buildup could occur. The practical lesson is that when building a forecasting model, if there isn't much difference in the goodness-of-fit statistics between the lagged and unlagged values of the independent variable, single-period forecast accuracy will almost always be improved by choosing lagged values of the independent variables.

5.6.3 Error Buildup in Multi-period Forecasting

A major source of error buildup in multi-period forecasting, in addition to the sources already mentioned, stems from using the lagged dependent variable on the right-hand side of the equation. Even if this term is not entered explicitly, the same general effect is generated any time an AR(1) transformation is used. Also, error buildup may occur from the predicted values of lagged independent variables. Even if all the independent variables in the equation are lagged, which means this source of error is absent for single-period forecasting, eventually these variables will become endogenous – and hence contain forecasting errors – if the forecast period is extended far enough in the future.

Foreign exchange rates are notoriously difficult to predict in any case, but illustrate the point well. The following case study utilizes an equation for the cross-rate between the Japanese yen and the US dollar. When the yen strengthens, it takes fewer yen to purchase one dollar, so $ ratio declines. From 1971, when the US went off the international gold standard, to early 1995, the yen rose from 360/$ to 80/$. It subsequently fell back to 140/$.

Explanations for the stronger yen during those years are plentiful: relative to the US, Japan had a lower rate of inflation, faster growth, and a big net export surplus. An equation estimated for 1971.1 through 1994.4 is examined next.

Case Study 9: The Yen/Dollar Cross-rate

The summary statistics for the yen/dollar equation estimated through 1994.4 are given in table 5.10. The key variables are the relative inflation rates, the Japanese growth rate, the US Federal funds rate, the change in the US rate of inflation, and the Japanese net export ratio. The residuals are shown in figure 5.10.

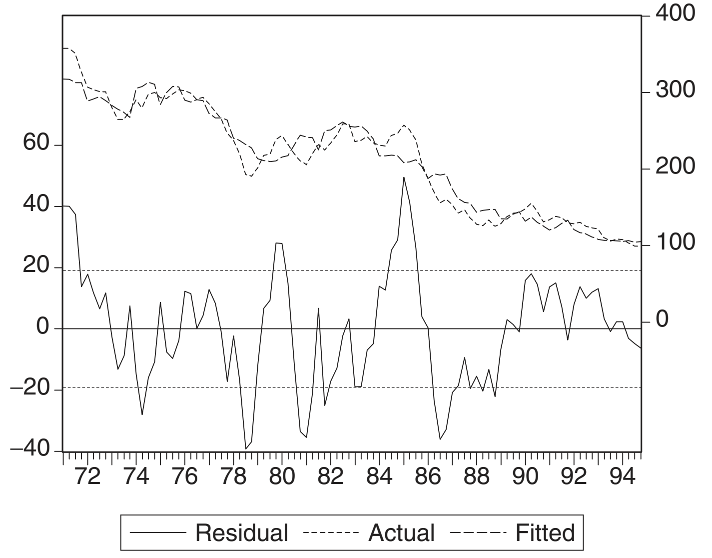

Figure 5.10 Residuals for the yen/dollar equation.

$$\begin{aligned} YEN &= 478.7 - \underset{(5.4)}{2.82} * (USINFL - JPNINFL) - \underset{(19.7)}{73.3} * JPNGDP \\ &\quad - \underset{(3.0)}{677} * JPNNEX - \underset{(3.9)}{3.17} * \Delta USINFL + \underset{(3.8)}{3.04} * RFF(-4) \\ &\quad RSQ = 0.928;\ DW = 0.44 \end{aligned}$$

where *YEN* is the yen/dollar cross-rate, *USINFL* is the US inflation rate, *JPNINFL* is the Japanese inflation rate, *JPNGDP* is Japanese real GDP, *JPNNEX* is Japanese ratio of net exports to GDP, and *RFF* is the Federal funds rate minus annual rate of inflation.

This equation fails to capture the spike in the dollar in 1985, but economists generally agree that was a speculative bubble unrelated to underlying economic forces; an offsetting opposite reaction occurred in 1986 after the G-7 took concerted action to reduce the value of the dollar. In recent years, the equation appears to be tracking the continued increase in the yen. At the end of 1994, one might have been pardoned for predicting that the yen would continue to appreciate indefinitely.

The other problem with this equation is the very high degree of autocorrelation, so the equation is rerun with AR(1). Ex post forecasts for the 1995.1–1998.4 period with (i) the original equation, (ii) the equation with AR(1), and (iii) in-sample estimates where the equation is reestimated through 1998.4 are presented in table 5.10.

Table 5.10 Residuals from alternative formulations of the yen/dollar equation.

Year/qtr	Actual yen/$	Equation 1		Equation 2		Equation 3	
		Predicted	Error	Predicted	Error	Predicted	Error
1995:1	96.2	99.6	–3.5	98.9	–2.8	101.4	–5.2
1995:2	84.5	101.1	–16.6	100.7	–16.2	103.1	–18.6
1995:3	94.2	105.9	–11.7	103.8	–9.5	108.1	–13.8
1995:4	101.5	108.5	–7.0	105.6	–4.1	111.0	–9.4
1996:1	105.8	98.9	6.9	103.5	2.3	102.1	3.8
1996:2	107.5	100.0	7.5	106.3	1.2	103.4	4.1
1996:3	109.0	100.1	8.9	108.6	0.4	103.7	5.3
1996:4	112.9	99.6	13.3	110.9	2.0	103.4	9.5
1997:1	121.2	106.4	14.9	113.8	7.4	110.2	11.0
1997:2	119.7	109.8	9.9	115.8	3.9	113.8	6.0
1997:3	118.1	110.7	7.4	116.9	1.1	114.7	3.4
1997:4	125.4	112.8	12.5	118.5	6.9	116.8	8.6
1998:1	128.2	120.4	7.8	121.0	7.2	123.7	4.4
1998:2	135.7	118.8	16.8	119.1	16.6	122.1	13.5
1998:3	140.0	118.0	22.0	117.9	22.1	121.1	18.9
1998:4	119.5	116.8	2.7	116.4	3.1	119.8	–0.3

None of the equations generate adequate forecasts. The equation with the AR(1) transform shows a gradually increasing trend in the yen/dollar ratio, while the original equation fails to track the turnaround in the yen until early 1997, whereas it actually turned up in mid-1995. An equation estimated through 1998 captures part of this turnaround but fails to predict the weakness in the yen in mid-1998 – a pattern that was again reversed in 1999. However, most of the forecasts in the AR(1) equation show a steadily increasing trend that fails to capture any of the turning points. Furthermore, this problem becomes more severe as the distance from the last sample point increases. Thus even though the equation appears to be robust from a theoretical viewpoint, the forecasts are not satisfactory. Adding the consensus forecast would not reduce forecast error either, since consensus forecasts of foreign exchange rates are almost random.

I stated at the outset of this book that it would occasionally be useful to show the failures as well as the successes of forecasting models. Perhaps this case study may seem to be carrying that principal to extremes, yet the following lessons can be gleaned.

1 Reestimating the equation with the AR(1) does not improve forecast accuracy. This point has already been emphasized several times.
2 This equation can be used to indicate the underlying value of the yen, and show when the actual rate diverges from equilibrium. Hence, for example, if the