view trading as a dynamic rather than static process between entry and exit points. The basic idea is that as a trade moves in the intended direction, the position exposure would be gradually reduced. The larger the move and the closer the market gets to a target objective, the more the position would be decreased. After reducing exposure in this manner, the position would be reinstated on a market correction. Any time the market retraced to a correction reentry point, a net profit would be generated that otherwise would not have been realized. The choppier the market, the more excess profits trading around the position will generate. Even a trade in which the market fails to move in the intended direction, on balance, could still be net profitable as a result of gains generated by lightening the total position on favorable trend moves and reinstating liquidated portions of the position on corrections. This strategy will also reduce the chances of being knocked out of a favorable position on a market correction, because if the position has already been reduced, the correction will have less impact and may even be desired to reinstate the liquidated portion of the position. The only time this strategy will have a net adverse impact is if the market keeps going in the intended direction without ever retracing to correction reentry levels. This negative outcome, however, simply means that the original trade was profitable, but the total profits are smaller than they would have been otherwise. In a nutshell, trading around a position will generate extra profits and increase the chances of staying with a good trade the at expense of sometimes giving up a portion of profits when the market moves smoothly in the intended direction.

27. **Being right is more important than being a genius.** I think one reason why so many people try to pick tops and bottoms is that they want to prove to the world how smart they are. Think about winning rather than being a hero. Forget trying to judge trading success by how close you can come to picking major tops and bottoms, but rather by how well you can pick individual trades with favorable return/risk characteristics. Go for consistency on a trade-to-trade basis, not perfect trades.

28. **Don't worry about looking stupid.** Last week, you told everyone at the office, "My analysis has just given me a great buy signal in the S&P. The market is going to a new high." Now as you examine the market action since then, something appears to be wrong. Instead of rallying, the market is breaking down. Your gut tells you that the market is vulnerable. Whether you realize it or not, your announced prognostications are going to color your objectivity. Why? Because you don't want to look stupid after telling the world that the market was going to a new high. Consequently, you are likely to view the market's action in the most favorable light possible. "The market isn't breaking down, it's just a pullback to knock out the weak longs." As a result of this type of rationalization, you end up holding a losing position far too long. There is an easy solution to this problem: Don't talk about your position.

What if your job requires talking about your market opinions (as mine once did)? Here the rule is: Whenever you start worrying about contradicting your previous opinion, view that concern as reinforcement to reverse your market stance. As a personal example, in early 1991, I came to the conclusion that the dollar had formed a major bottom. I specifically remember one talk in which an audience member asked me about my outlook for currencies. I responded by boldly predicting that the dollar would head higher for years. Several months later, when the dollar surrendered the entire gain it had realized following the news of the August 1991 Soviet coup before the coup's failure was confirmed, I sensed that something was wrong. I recalled my many predictions over the preceding months in which I had stated that the dollar would go up for years. The discomfort and embarrassment I felt about these previous forecasts told me it was time to change my opinion.

In my earlier years in the business, I invariably tried to rationalize my original market opinion in such situations. I was burned enough times so that I eventually learned a lesson. In the preceding example, the abandonment of my original projection was fortunate because the dollar collapsed in the ensuing months.

29. **Sometimes action is more important than prudence.** Waiting for a price correction to enter the market may sound prudent, but it is often the wrong thing to do. When your analysis, methodology, or gut tells you to get into a trade at the market instead of waiting for a correction—do so. Caution against the influence of knowing that you could have gotten in at a better price in recent sessions, particularly in those situations when the market witnesses a sudden, large move (often due to an important surprise news item). These types of trades often work because they are so hard to do.
30. **Catching part of the move is just fine.** Just because you missed the first major portion of a new trend, don't let that keep you from trading with that trend (as long as you can define a reasonable stop-loss point). McKay commented that the easiest part of a trend is the middle portion, which implies always missing part of the trend prior to entry.
31. **Don't try to be 100 percent right.** Almost every trader has had the experience of the market moving against the position sufficiently to raise significant concern regarding the potential additional loss, while still believing the position is correct. Staying in the trade risks an uncomfortably large loss, but liquidating the trade risks abandoning a good position at nearly the worst possible point. In such circumstances, instead of making an all-or-nothing decision, traders can choose to liquidate part of the position. Taking a partial loss is much easier than liquidating the entire position and will avoid the possibility of riding the entire position for a large loss. It will also preserve the potential for a partial recovery if the market turns around.
32. **Maximize gains, not the number of wins.** Eckhardt explains that human nature does not operate to maximize gain but rather the chance of a gain. The problem with this is that it implies a lack of focus on the magnitudes of gains (and losses)—a flaw that leads to nonoptimal performance results. Eckhardt bluntly concludes: "The success rate of trades is the least important performance statistic and may even be inversely related to performance." Jeff Yass, a very successful options trader, echoes a similar theme: "The basic concept that applies to both poker and option trading is that the primary object is not winning the most hands, but rather maximizing your gains."
33. **Learn to be disloyal.** Loyalty may be a virtue in family, friends, and pets, but it is a fatal flaw for a trader. Never have loyalty to a position. The novice trader will have lots of loyalty to his original position. He will ignore signs that he is on the wrong side of the market, riding his trade into a large loss while hoping for the best. The more experienced trader, having learned the importance of money management, will exit quickly once it is apparent he has made a bad

trade. However, the truly skilled trader will be able to do a 180-degree turn, *reversing* his position at a loss if market behavior points to such a course of action. Druckenmiller made the awful error of reversing his stock position from short to long on the very day before the October 19, 1987, crash. His ability to quickly recognize his error and, more important, to unhesitatingly act on that realization by reversing back to short at a large loss helped transform a potentially disastrous month into a net profitable one.

34. **Pull out partial profits.** Pull a portion of winnings out of the market to prevent trading discipline from deteriorating into complacency. It is far too easy to rationalize overtrading and procrastination in liquidating losing trades by saying, "It's only profits." Profits withdrawn from an account are much more likely to be viewed as real money.
35. **Hope is a four-letter word.** Hope is a dirty word for a trader, not only in regards to procrastinating in a losing position, hoping the market will come back, but also in terms of hoping for a reaction that will allow for a better entry in a missed trade. If such trades are good, the hoped-for reaction will not materialize until it is too late. Often, the only way to enter such trades is to do so as soon as a reasonable stop-loss point can be identified.
36. **Don't do the comfortable thing.** Eckhardt offers the rather provocative proposition that the human tendency to select comfortable choices will lead most people to experience worse than random results. In effect, he is saying that natural human traits lead to such poor trading decisions that most people would be better off flipping coins or throwing darts. Some of the examples Eckhardt cites of the comfortable choices people tend to make that run counter to sound trading principles include gambling with losses, locking in sure winners, selling on strength and buying on weakness, and designing (or buying) trading systems that have been overfitted to past price behavior. The implied message to the trader is: do what is right, not what feels comfortable.
37. **You can't win if you have to win.** There is an old Wall Street adage: "Scared money never wins." The reason is quite simple: If you are risking money you can't afford to lose, all the emotional pitfalls of trading will be magnified. Early in his career, when the bankruptcy of a key financial backer threatened the survival of his fledgling investment firm, Druckenmiller "bet the ranch" on one trade, in a last-ditch effort to save his firm. Even though he came within one week of picking the absolute bottom in the T-bill market, he still lost all his money. The need to win fosters trading errors (e.g., excessive leverage and a lack of planning in the example just cited). The market seldom tolerates the carelessness associated with trades born of desperation.
38. **The road to success is paved with mistakes.** Learning from mistakes is essential to improvement and ultimate success. Each mistake, if recognized and acted on, provides an opportunity for improving a trading approach. Most traders would benefit by writing down each mistake, the important lesson, and the intended change in the trading process. Such a trading log can be periodically reviewed for reinforcement. Trading mistakes cannot be avoided, but repeating the same mistakes can be, and doing so is often the difference between success and failure.
39. **Think twice when the market lets you off the hook easily.** Don't be too eager to get out of a position you have been worried about if the market allows you to exit at a much better price than anticipated. If you had been worried about an adverse overnight (or over-the-weekend)

price move because of a news event or a technical price failure on the previous close, it is likely that many other traders shared this concern. The fact that the market does not follow through much on these fears strongly suggests that there must be some very powerful underlying forces in favor of the direction of the original position. This concept, which was first proposed in *Market Wizards* by Marty Schwartz, who compiled an astounding track record trading stock index futures, was illustrated by the manner in which Lipschutz, a large-scale currency trader, exited the one trade he admitted had scared him. In that instance, on Friday afternoon, a time when the currency markets are particularly thin (after Europe's close), Lipschutz found himself with an enormous short dollar position in the midst of a strongly rallying market. He had to wait over the weekend for the Tokyo opening on Sunday evening to find sufficient liquidity to exit his position. When the dollar opened weaker than expected in Tokyo, he didn't just dump his position in relief; rather, his trader's instincts told him to delay liquidation—a decision that resulted in a far better exit price.

40. **A mind is a terrible thing to close.** Open-mindedness seems to be a common trait among those who excel at trading. For example, Gil Blake, a mutual fund timer who has made incredibly consistent profits, actually fell into a trading career by attempting to demonstrate to a colleague that prices were random. When he realized he was wrong, he became a trader. In the words of Driehaus, "The mind is like a parachute—it's only good when it's open."
41. **The markets are an expensive place to look for excitement.** Excitement has a lot to do with the image of trading, but nothing to do with success in trading (except in an inverse sense). In *Market Wizards,* Larry Hite, the founder of Mint Management, one of the largest CTA firms, described his conversation with a friend who couldn't understand his absolute adherence to a computerized trading system. His friend asked, "Larry, how can you trade the way you do? Isn't it boring?" Larry replied, "I don't trade for excitement; I trade to win."
42. **Beware of trades born of euphoria.** Take caution against placing impulsive trades influenced by being caught up in market hysteria. Excessive euphoria in the market should be seen as a cautionary flag of a potential impending reversal.
43. **If you are on the right side of euphoria or panic, lighten up.** Parabolic price moves tend to end abruptly and sharply. If you are fortunate enough to be on the right side of the market in which the price move turns near vertical, consider scaling out of the position while the trend is still moving in your direction. If you would be petrified to be on the other side of the market, that is probably a good sign that you should be lightening your position.
44. **The calm state of a trader.** If there is an emotional state associated with successful trading, it is the antithesis of excitement. Based on his observations, Charles Faulkner, a neuro-linguistic programming (NLP) practitioner who works with traders, stated that exceptional traders are able to remain calm and detached regardless of what the markets are doing. He describes Peter Steidlmayer's (a successful futures trader who is best known as the inventor of the Market Profile trading technique) response to a position that is going against him as being typified by the thought, "Hmmm, look at that."
45. **Identify and eliminate stress.** Stress in trading is a sign that something is wrong. If you feel stress, think about the cause, and then act to eliminate the problem. For example, let's say you

determine that the greatest source of stress is indecision in getting out of a losing position. One way to solve this problem is simply to enter a protective stop order every time you put in a position.

I will give you a personal example. When I was a research director, one of the elements of my job was providing trading recommendations to brokers in my company. This task is very similar to trading, and, having done both, I believe it's actually more difficult than trading. At one point, after years of net profitable recommendations, I hit a bad streak. I just couldn't do anything right. When I was right about the direction of the market, my buy recommendation was just a bit too low (or my sell price too high). When I got in and the direction was right, I got stopped out—frequently within a few ticks of the extreme of the reaction.

I responded by developing a range of computerized trading programs and technical indicators, thereby widely diversifying the trading advice I provided to the firm. I still made my day-to-day subjective calls on the market, but everything was no longer riding on the accuracy of these recommendations. By widely diversifying the trading-related advice and information, and transferring much of this load to mechanical approaches, I was able to greatly diminish a source of personal stress—and improve the quality of the research product in the process.

46. **Pay attention to intuition.** As I see it, intuition is simply experience that resides in the subconscious mind. The objectivity of the market analysis done by the conscious mind can be compromised by all sorts of extraneous considerations (e.g., one's current market position, a resistance to change a previous forecast). The subconscious, however, is not inhibited by such constraints. Unfortunately, we can't readily tap into our subconscious thoughts. However, when they come through as intuition, the trader needs to pay attention. As the Zen-quoting trader mentioned earlier expressed it, "The trick is to differentiate between what you *want* to happen and what you *know* will happen."
47. **Life's mission and love of the endeavor.** In talking to the traders interviewed in *Market Wizards,* I had the definite sense that many of them felt that trading was what they were meant to do—in essence, their mission in life. In this context, Charles Faulkner quoted NLP cofounder John Grinder's description of mission: "What do you love so much that you would pay to do it?" Throughout my interviews, I was struck by the exuberance and love the Market Wizards had for trading. Many used gamelike analogies to describe trading. This type of love for the endeavor may indeed be an essential element for success.
48. **The elements of achievement.** Faulkner has a list of six key steps to achievement based on Gary Faris's study of successfully rehabilitated athletes, which appears to apply equally well to the goal of achieving trading success. These strategies include the following:
    1. Using both "Toward" and "Away From" motivation;
    2. Having a goal of full capability plus, with anything less being unacceptable;
    3. Breaking down potentially overwhelming goals into chunks, with satisfaction garnered from the completion of each individual step;
    4. Keeping full concentration on the present moment—that is, the single task at hand rather than the long-term goal;

5. Being personally involved in achieving goals (as opposed to depending on others); and
6. Making self-to-self comparisons to measure progress.

49. **Prices are nonrandom = the markets can be beat.** In reference to academicians who believe market prices are random, Monroe Trout, a commodity trading advisor with one of the best risk/return records in the industry, says, "That's probably why they're professors and why I'm making money doing what I'm doing." The debate over whether prices are random is not yet over. However, my experience in interviewing scores of great traders left me with little doubt that the random walk theory is wrong. It is not the magnitude of the winnings registered by the Market Wizards, but the consistency of these winnings in some cases, that underpin my belief. As a particularly compelling example, in his first fund, Edward Thorp, a mathematician best known for his best-selling book *Beat the Dealer*, compiled a track record of 227 winning months and only 3 losing months (all under 1 percent)—an extraordinary 98.7 winning percentage. The odds of getting such a result by chance (as would be the case if the markets were random) are less than 1 out of 10^{63}. To put this probability in context, the odds of randomly selecting a specific atom in the earth would be about a trillion times better. Certainly, winning at the markets is not easy—and, in fact, it is getting more difficult as professionals account for a constantly growing proportion of the activity—but it can be done!

50. **Keep trading in perspective.** There is more to life than trading.

APPENDIX A

Introduction to Regression Analysis

Theory helps us bear our ignorance of fact.

—George Santayana

Basics

Regression analysis is concerned with describing and evaluating the relationship between a given variable and one or more other variables. For example, we might be interested in describing the relationship between the pig crop (number of pigs born during a given period) and the hog slaughter level in the following six-month period.[1] The relationship between these variables is illustrated in Figure A.1. Each point in Figure A.1 represents a single observation or year. The location of a point along the horizontal axis is determined by the December–May pig crop, while its placement along the vertical axis is determined by the June–November hog slaughter level. Note that there is a clear

[1] Readers may notice that a predominant number of the examples in the Appendices will be drawn from the hog market. There are three basic reasons for this: (1) Such comparisons will illustrate the advantages of regression analysis in terms of preciseness, efficiency, flexibility, and ease of application. (2) The exposition will be clearer if a limited number of markets are used to provide illustrative examples. (3) Because hogs are nonstorable, the hog market can be represented adequately by simple fundamental models. In any event, it should be stressed that chosen examples are merely intended as vehicles to illustrate the general concepts and techniques of regression analysis, and not as a description of the methodology for analyzing any specific market. Consequently, the illustrations should be as relevant to the reader interested in applying regression analysis to the interest rate markets as to the reader whose primary focus is the livestock sector.

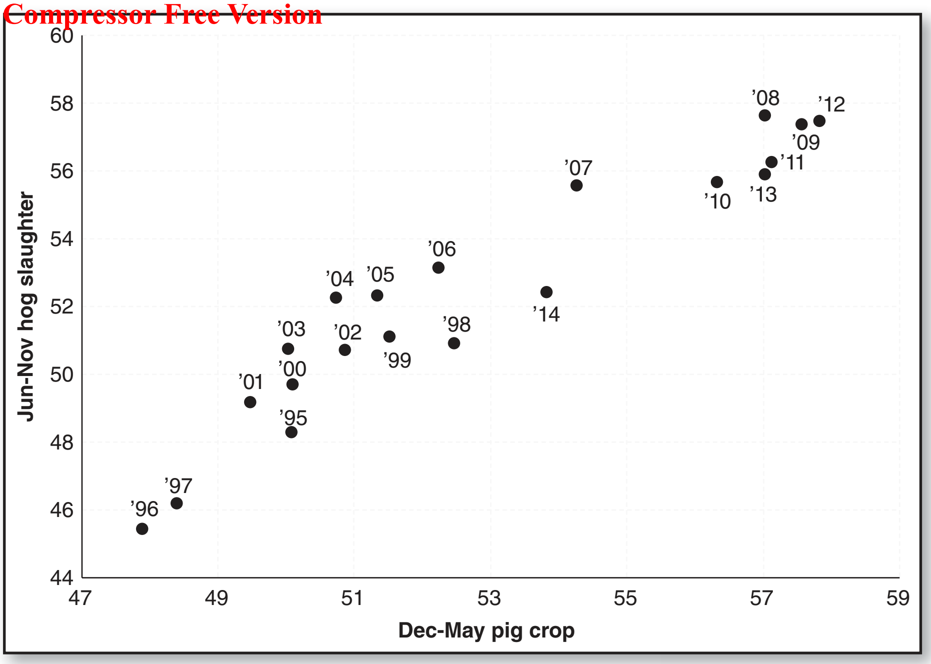


FIGURE A.1 June–November Hog Slaughter vs. December–May Pig Crop (Thousands)

relationship between these two variables: large hog slaughter levels correspond to large pig crop levels. In this example, hog slaughter is the *dependent* variable in that hog slaughter depends on the pig crop, but not vice versa, and the pig crop is the *independent*, or *explanatory*, variable. The primary goal of regression analysis is to define a mathematical relationship between the dependent variable and the independent variable(s).

Perhaps the most basic underlying assumption in the standard regression analysis approach is that the relationship between the dependent and independent variables is linear. In the case in which there is only one explanatory variable, the regression equation will be a straight line and can be expressed as

$$Y = a + bX$$

where a and b are constants determined by the regression procedure.[2] The values derived for a and b by the regression procedure are termed the *regression coefficients* (a is sometimes simply referred to as the *constant term*). By convention, Y is the variable that we are trying to explain or predict—the dependent variable—while X is the explanatory or independent variable.

[2] To be precise, a and b are parameters. A parameter can be thought of as a hybrid between a variable and a constant. If the focus is on the variation of the equation as a whole, then a and b are variables. Given the equation, $Y = a + bX$, each set of values for a and b will define a different line. However, if we are concerned with the relationship between the variables X and Y, given a specific set of values for a and b, as is the case in regression analysis, then a and b can be termed constants.

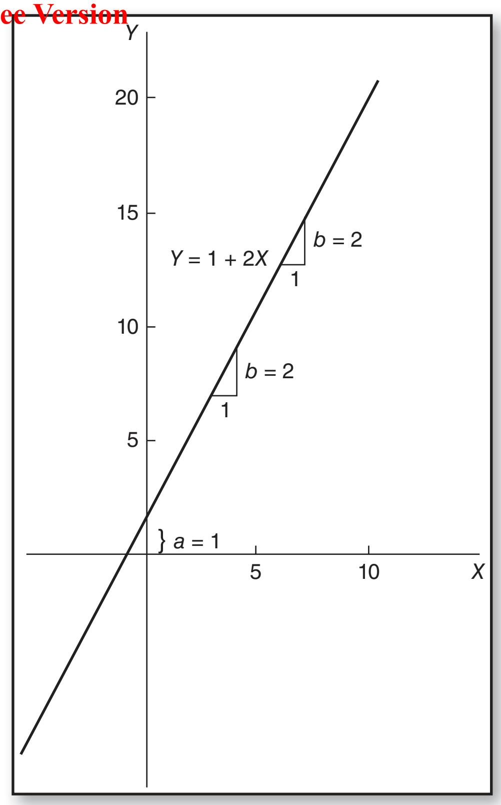


FIGURE A.2 Meaning of *a* and *b* for Straight Line

The constants *a* and *b* in the regression equation have special meanings. Constant *b* is the amount variable *Y* (e.g., hog slaughter) will change given a one-unit change in variable *X* (e.g., pig crop). For example, in the simple linear equation $Y = 1 + 2X$ each unit change in *X* will result in a two-unit change in *Y*. Note this relationship will hold regardless of the level of *X*. In fact, the constancy of the change in *Y* given a fixed change in *X* is a basic characteristic of a linear equation. Constant *a* is called the *Y* intercept because it is the value of *Y* at which the line crosses the *Y* axis—that is, the value of *Y* when *X* equals zero. (See Figure A.2 for a graphic depiction of the preceding points.)

Given a set of data points such as those illustrated in Figure A.1, regression analysis will seek to find the values of *a* and *b* in the regression equation that result in the line that best fits the observed points.

Meaning of Best Fit

Using Figure A.1 as an example, how would we define the best-fit line to the scatter of points? Intuitively, it seems that we would want to pick the line that minimizes the deviations from the individual points to the line. The *deviation* of any single point or observation can be defined as the difference

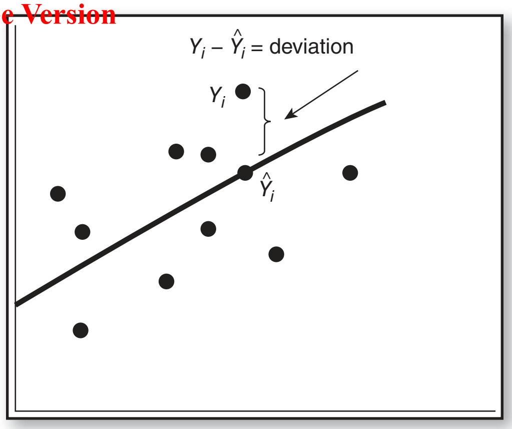


FIGURE A.3 Deviation for a Single Observation

between Y_i, the observed value, and $\hat{Y}_i$, the Y value predicted by the line for the same value of X. The deviation of a single point is thus equal to $Y_i - \hat{Y}_i$ (see Figure A.3).

These deviations are also called *residuals.* We cannot derive a summary deviation figure for a group of points by adding all the individual deviations. Why? Because deviations above and below the line will tend to cancel each other out. Thus, the sum of the residuals can be small even if the line fits the data points poorly. In fact, if the deviations below the line are greater than the deviations above the line, the sum of the residuals will be negative—an absurd value for a measure of total deviation. How would one interpret a negative total deviation? In other words, the sum of the residuals does not offer a criterion for determining best fit.

One possible solution is to find the line that minimizes the sum of the *absolute* deviations, that is, the sum of the residuals measured without regard to sign. Another possible approach would be to square each of the deviations before adding them, thereby assuring that they will all be positive, and then to find the line that minimizes the sum of these squared deviations[3]:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

This *least-squares* approach represents the method employed by regression analysis, and is preferable to the sum of the absolute deviations for several reasons:

1. Theoretically, the least-squares approach will yield the best estimates.[4]
2. The least-squares method will place greater weight on large errors as a result of the squaring operation in its computation. This approach is usually advantageous, since it is desirable to avoid large deviations.

[3] The symbol Σ means “the sum of.” The superscript n indicates the number of observations, and the subscript $i = 1$ indicates the observation number at which the summation begins. In other words, in this term, all the squared deviations are summed, and there are a total of n observations.

[4] The least-squares estimates will be both *unbiased* and *efficient*. These terms are defined in Appendix C.

3. The sum of the absolute deviations is computationally far more unwieldy than the sum of the squared deviations.
4. The least-squares approach permits many useful tests of the reliability of the equation.

It can be demonstrated by straightforward calculus proofs the values of *a* and *b* that minimize the sum of the squared deviations are:

$$b = \frac{n \cdot \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2}$$

$$a = \frac{\sum_{i=1}^{n} Y_i}{n} - b \frac{\sum_{i=1}^{n} X_i}{n} = \overline{Y} - b\overline{X}$$

where n = number of observations
$\overline{Y}$ = mean of Y_i, and
$\overline{X}$ = mean of X_i

A Practical Example

As a practical example, we will find the best-fit line using the least-squares approach for the set of observations in Figure A.1. Table A.1 summarizes the necessary computations. The resulting best-fit line is illustrated in Figure A.4. To obtain a specific forecast, we would merely plug the estimated pig crop value into the regression equation. For example, if the December–May pig crop estimate were 51 million, the forecast for hog slaughter in the subsequent June–November period would be 50.51 million (–3.6279 + (1.0615 * 51)).

Reliability of the Regression Forecast

It is essential to understand that, by itself, a point price projection derived from a regression equation is of little use. One must first consider how well the model describes the data and the expected variability of forecasts based upon the regression equation. We can get an intuitive answer to this question by examining how closely the observations fall to the fitted regression line (Figure A.4).

But we should be able to assess a model's accuracy more precisely. Simply examining a scatter chart leaves many unanswered questions. How close do the observations have to be to the regression line for the model to be judged satisfactory? How do we check whether a model provides an undistorted representation of the real world? How closely can we expect the model's forecasts to anticipate actual results?

TABLE A.1 Computation of Least-Squares Best-Fit Line

Year	Pig Crop (Dec–May, millions) X_i	Hog Slaughter (Jun–Nov, millions) Y_i	X_i^2	X_iY_i
1995	50.077	48.294	2,507.71	2,418.40
1996	47.888	45.453	2,293.26	2,176.64
1997	48.394	46.201	2,341.98	2,235.85
1998	52.469	50.929	2,753.00	2,672.20
1999	51.519	51.111	2,654.21	2,633.20
2000	50.087	49.689	2,508.71	2,488.76
2001	49.472	49.169	2,447.48	2,432.50
2002	50.858	50.709	2,586.54	2,578.94
2003	50.029	50.758	2,502.90	2,539.38
2004	50.737	52.265	2,574.24	2,651.76
2005	51.33	52.333	2,634.77	2,686.23
2006	52.242	53.150	2,729.23	2,776.68
2007	54.266	55.569	2,944.80	3,015.52
2008	57.019	57.648	3,251.17	3,287.05
2009	57.564	57.391	3,313.61	3,303.68
2010	56.326	55.681	3,172.62	3,136.26
2011	57.118	56.264	3,262.47	3,213.69
2012	57.818	57.478	3,342.92	3,323.23
2013	57.02	55.914	3,251.28	3,188.23
2014	53.821	52.418	2,896.70	2,821.17
	$\Sigma X_i = 1{,}056.05$	$\Sigma Y_i = 1{,}048.42$	$\Sigma X_i^2 = 55{,}969.58$	$\Sigma X_iY_i = 55{,}579.37$

b = (20 * 55,579.37) – (1,056.05 * 1,048.42) / (20 * 55,969.58) – $(55{,}969.58)^2$ = 1.0615

a = (1,048.42/20) – 1.0615 * (1,056.05/20) = –3.6279

$Y_i = -3.6279 + 1.0615X_i$

Another problem with the graphic analysis depicted in Figure A.4 is that it just isn't feasible for regression equations that include two or more explanatory variables—a situation that is the rule rather than the exception.

These considerations lead us to one of the primary benefits of regression analysis: The approach permits a wide variety of scientific tests of a model's adequacy. Such tests are essential to the successful application of regression analysis. An understanding of these tests, as opposed to a mere cookbook application, requires a synopsis of some key statistical concepts. Appendix B provides an abridged crash course in elementary statistics. We will return to regression analysis in Appendix C.

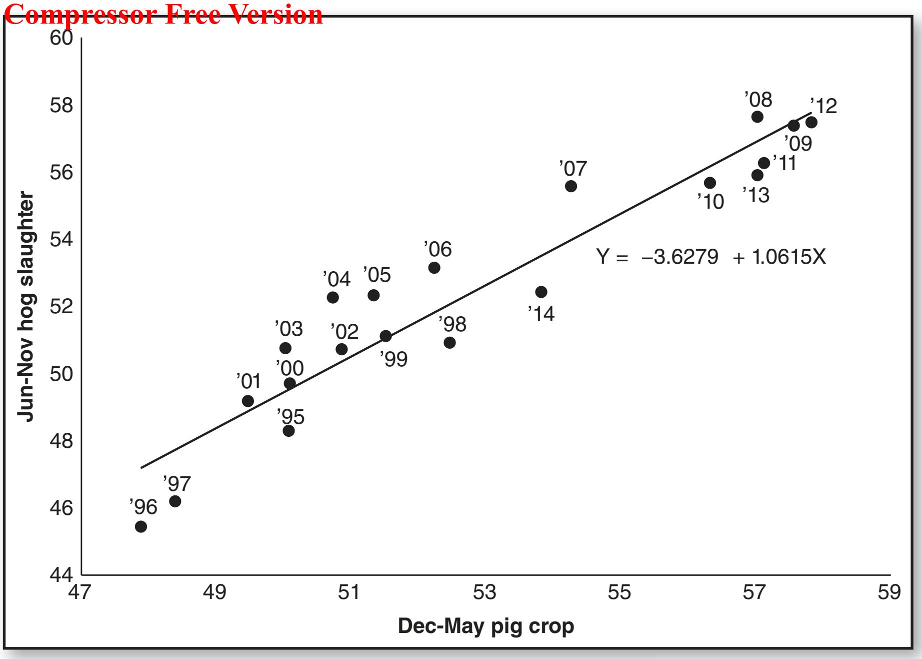


FIGURE A.4 Best-Fit Line for June–November Hog Slaughter vs. December–May Pig Crop

A Review of Elementary Statistics

The theory of probabilities is at bottom nothing but common sense reduced to Calculus.

—Pierre Simon de Laplace

Measures of Dispersion

For any data series there are two basic types of descriptive statistics: (1) some measure of central tendency (e.g., arithmetic mean, median, mode, geometric mean, harmonic mean); and (2) a measure of dispersion. The intuitive meaning of dispersion is quite clear. For example, consider the following two sets of numbers:

A. 30, 53, 3, 22, 16, 104, 71, 41
B. 42, 40, 42, 46, 39, 45, 42, 44

Although both series have the same arithmetic mean, it is clear that series A would have a high dispersion measure and series B a low dispersion measure. The concept of dispersion is extremely important in forecasting. For example, if we were told there was a ninth number in each of the series that was not listed, we would be far more certain about our guess being close to the mark in series B than in series A. Thus, it is extremely desirable to have a measure that describes the dispersion of a set of numbers, much as the mean describes the central tendency of a set of numbers.

The basic question is: How do we measure dispersion? In a sense, we have already answered this question. Deriving a dispersion measure for a set of numbers is entirely analogous to the computation of a single deviation measure for a group of points from a line. In the case of a set of numbers, the deviations would be measured relative to some central point. For theoretical reasons, the arithmetic mean is the most desirable measure of central tendency. To derive a single deviation measure for a set of numbers, we cannot simply add the individual deviations, because they will tend to cancel each other out. Once again, two possible solutions are the sum of the absolute deviations or the sum of the squared deviations. The latter measure is far more convenient to use and is preferable for theoretical reasons.

However, the sum of the squared deviations is not a representative measure of dispersion since it is dependent on how many numbers are in the series. For example, if series B contained 1,000 sets of the indicated string of numbers, the sum of the squared deviations for the series would be greater than the corresponding figure for series A. This measure is therefore quite misleading because series A would still reflect greater dispersion by any intuitive definition of that term. This problem is solved simply by dividing the sum of the squared deviations by the number of items in the series. The resulting measure is called the *variance,* which can be expressed as:

$$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}$$

where $\overline{X}$ = mean
X_i = individual data values
N = number of observations

Note the variance is not stated in the same units as the original data series. For example, if the units of the original set of numbers were tons, the variance would be expressed in tons squared.

The dispersion measure can be expressed in the same units as the original data series by simply taking the square root of the variance. This computation also makes intuitive sense since it reverses the original squaring process applied to the individual terms. The resulting figure is called the *standard deviation* and can be expressed as:

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}}$$

In a rough sense, the standard deviation is a type of average deviation (of the individual data points from the mean), in which the data points that are further from the mean have greater than proportionate impact on the calculation. (This greater weight is the result of the squaring process.)[1]

[1] These definitions for the variance and standard deviation are applicable when the entire set of data elements is known, in which case the set of numbers is called the *population.* However, in actual practice, available sets of numbers will often represent *samples* from a population. In fact, this assumption appears to be implied for series A and B. For reasons that will be explained later, the variance and standard deviation calculations for a sample are slightly different. Specifically, for samples, the variance and standard deviation would be expressed as follows:

$$\text{Variance (sample)} = s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}$$

$$\text{Standard deviation (sample)} = s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}}$$

where n = number of observations in the sample.

TABLE B.1 Standard Deviation Computations

Series A: 30, 53, 3, 22, 16, 104, 71, 41			Series B: 42, 40, 42, 46, 39, 45, 42, 44		
X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
30	−12.5	156.25	42	−0.5	0.25
53	+10.5	110.25	40	−2.5	6.25
3	−39.5	1,560.25	42	−0.5	0.25
22	−20.5	420.25	46	+3.5	12.25
16	−26.5	702.25	39	−3.5	12.25
104	+61.5	3,782.25	45	+2.5	6.25
71	+28.5	812.25	42	−0.5	0.25
41	−1.5	2.25	44	+1.5	2.25
$\sum_{i=1}^{n} X_i = 340$		$\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 = 7,546.00$	$\sum_{i=1}^{N} X_i = 340$		$\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 = 40.00$
$\bar{X} = \frac{\sum X_i}{N} = 42.5$			$\bar{X} = \frac{\sum X_i}{N} = 42.5$		
$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{N} = \frac{7,546}{8} = 943.25$			$\text{Variance} = \sigma^2 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{N} = \frac{40}{8} = 5$		
$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{N}} = 30.712$			$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{N}} = 2.236$		

Note: These computations apply to a population. For samples, the computation would be slightly different (see footnote 1).

The greater the standard deviation, the greater the degree of variability in a set of numbers. To get a better sense of this statistic, Table B.1 calculates the standard deviation for series A and B. It is essential to have a clear understanding of the standard deviation before proceeding, since this term will play a pivotal role in defining the *normal distribution* and in probability testing.

Probability Distributions

A *random variable* is a variable with a value that depends on a statistical experiment in which each outcome (or range of outcomes) has a specific probability of occurrence. For example, if trading decisions were based on the toss of a coin, the number of winning trades, excluding commissions, in 10 trades would be a random variable. A *probability distribution* indicates the probability associated with different values of a random variable. Figure B.1 indicates the probabilities for different numbers of gains in 10 trades if trading decisions are based on chance. The highest probability of 0.246 is associated with five gains in 10 trades. The probability of alternative events decreases as the number of gains moves

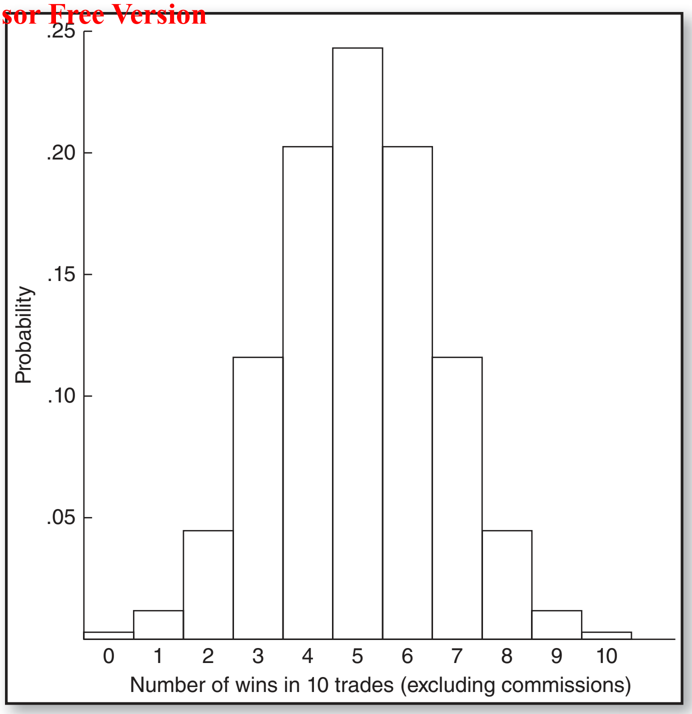


FIGURE B.1 Probability Distribution for Number of Winning Trades in 10 Trades If Decision Based on Chance

away from five. The probability of 10 out of 10 winning trades is only 0.001. (By definition, the sum of all the probabilities equals 1.0.)

This example of a probability distribution was based on a *discrete* variable, which is a variable that can take on only certain fixed values—for example, we can have six winning trades or seven winning trades, but not 6.3 winning trades. Frequently, we will be concerned with random variables that are *continuous,* which are variables that can assume any value. An example of a continuous variable would be the reaction time of drivers in stepping on the brake when a stop sign is flashed on a screen in a simulation test. For continuous variables, the probability of each event (e.g., probability of the reaction time being exactly 0.41237 second) is not meaningful or even definable. Instead, the relevant consideration is the probability of events in a certain range (e.g., the probability of a reaction time between 0.4 and 0.5 seconds).

A *continuous distribution* describes the probability associated with a continuous random variable. The total area under a continuous distribution curve will equal 1.0 (100 percent) since there is 100 percent probability an event will take on some value, and the sum of all the probabilities of mutually exclusive events cannot exceed 100 percent.[2] A continuous distribution is characterized by the

[2] *Mutually exclusive* means that only one event can occur at a time. For example, in the reaction time test, only one time value can be associated with any given test.

fact that the area between any two given values is equal to the probability the random variable will fall in the interval between these two values. For example, in Figure B.2 the total area under the curve would be equal to 1.0, and the shaded area would indicate the probability of the continuous variable having a value between X_1 and X_2. If the shaded area represented 20 percent of the total area under the curve, the probability of the continuous variable falling in a range between X_1 and X_2 would be 20 percent.

Figure B.2 represents the familiar bell-shaped *normal distribution* curve. Empirically, the normal distribution has been shown to serve as a good approximation of the probability distribution for an extremely wide range of random variables. For example, it can be demonstrated that as the number of trades in Figure B.1 increases, the distribution will begin to approach a normal distribution. For a large number of trades (e.g., 1,000), the probability distribution would be almost exactly represented by a normal distribution. Probabilities for continuous random variables such as reaction time frequently will also be well described by the normal distribution.

Figure B.3 shows how the probability of an event falling within a fixed interval increases as the interval moves closer to the mean. The probability of an event occurring in the range $X_1 - X_2$ (i.e., the area under the curve between X_1 and X_2) is greater than the probability of an event in the range

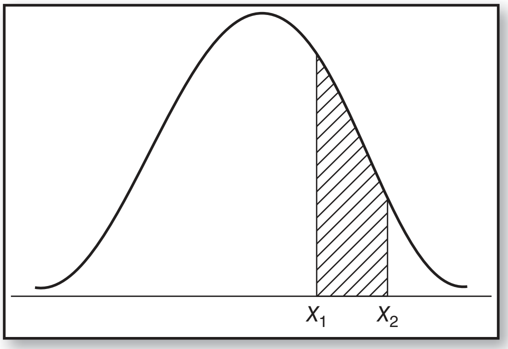


FIGURE B.2 Continuous Probability Distribution

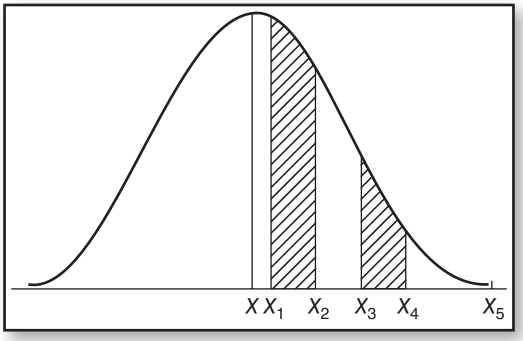


FIGURE B.3 Fixed Interval Probability Increases with Proximity to Mean

X_3–X_4. Note the probability of an event occurring in a range distant from the mean is near zero, even if it is a very broad range. For example, in Figure B.3, the probability of the variable having a value between X_5 and infinity is near zero.

The formula for the normal distribution is:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)\left[(X-\bar{X})/\sigma\right]^2}$$

This seemingly intimidating formula is not as frightening as it might initially appear. Like any other equation describing a relationship between X and Y, it tells us the value of Y given a value for X. The key point to realize about this equation is the precise relationship between X and Y will be determined entirely by the mean of $X(\bar{X})$ and the variance of X (σ).[3] All the other values in the formula are constants ($\pi = 3.1416$, $e = 2.7183$). Thus, once $\bar{X}$ and σ are determined, the normal distribution for a particular set of numbers is completely defined. Note the value of Y will reach a maximum when X equals $\bar{X}$, at which point the formula reduces to

$$Y = \frac{1}{\sigma\sqrt{2\pi}}$$

At any other value of X, the value of the term

$$\frac{1}{2}\left(\frac{X - \bar{X}}{\sigma}\right)^2$$

will be greater than 0, resulting in a lower value of Y. The further any given value X is from $\bar{X}$, the larger this term and the lower the value of Y.[4]

Because the normal distribution will differ for any given set of values for $\bar{X}$ and σ, it is desirable to choose a given set of values upon which to base a standard table of probability values. For simplicity, this table is based on $\bar{X} = 0$ and $\sigma = 1$. To be able to use this standard table, we have to transform the numbers in a series into Z values, where

$$Z_i = \frac{X_i - \bar{X}}{\sigma_x}$$

[3] $\bar{X}$ and σ are parameters. As explained in footnote 2 in Appendix A, a parameter can be thought of as a hybrid between a variable and a constant. In this instance, $\bar{X}$ and σ will assume different values for different distributions of X (i.e., different sets of numbers); however, for any given distribution (set of numbers), $\bar{X}$ and σ will be fixed (i.e., constants).

[4] e^{-k} is equivalent to $1/e^k$, therefore the larger $1/2[(X - \bar{X})/\sigma]^2$ gets, the smaller the value of $e^{-(1/2)[(X-\bar{X})/\sigma]^2}$, hence the smaller the value of Y.

and X_i is a given value in a set of numbers.[5] The numerator of this term is the distance of the given number from the mean; the denominator is the standard deviation of the set of numbers. Thus, the Z value is simply the distance of a given value from the mean in terms of standard deviations. For example, if the mean of a set of numbers is 10, and the standard deviation is 2, the Z value for a

[5] The fact that the distribution of Z values will always have a mean equal to zero ($\bar{Z} = 0$) and a standard deviation equal to 1 ($\sigma_z = 1$) given that any set of X values is easy to demonstrate:

$$Z = \frac{X_i - \bar{X}}{\sigma_X} \qquad \bar{Z} = \frac{\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{\sigma_X}\right)}{N} = \frac{\frac{1}{\sigma_X}\left(\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \bar{X}\right)}{N}$$

Keeping in mind that $\bar{X} = \left(\sum_{i=1}^{N} X_i\right) / N$.

$$\bar{Z} = \frac{1}{N\sigma_X}\left(N\bar{X} - N\bar{X}\right) = 0$$

The standard deviation of Z (σ_z) can be expressed as

$$\sigma_z = \sqrt{\frac{\sum_{i=1}^{N}\left(Z_i - \bar{Z}\right)^2}{N}}$$

But we have just proved that $\bar{Z} = 0$, so

$$\sigma_Z = \sqrt{\frac{\sum_{i=1}^{N} Z_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2}{N}} = \sqrt{\frac{1}{\sigma_x^2} \cdot \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{N}}$$

$$\sigma_Z = \frac{1}{\sigma_X}\sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{N}}$$

Since

$$\sqrt{\frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}{N}}$$

is the definition for σ_x,

$$\sigma_Z = \frac{1}{\sigma_X} \cdot \sigma_X = 1$$

number $X = 6$ would be -2 (i.e., X is 2 standard deviations removed from the mean). This standardized distance of a number from its mean will allow us to gauge the probabilities of a given value being higher or lower than a given number.

Reading the Normal Curve (*Z*) Table

Remember, a Z value indicates how many standard deviations a given observation lies above or below its mean, with the sign indicating whether the number is above or below the mean. Table B.2 lists the probabilities corresponding to different Z values. These numbers represent the probabilities of an observation of a normally distributed random variable falling in the range between zero and the given Z value. For example, there is a .4332 (43.32 percent) probability the Z value will be between zero and $+1.5$. To determine the probability of a Z value being less than a given number, simply add .50 (the probability of a value below the mean) to the probability listed in Table B.2. Thus, the probability of a Z value less than $1.5 = .9332$. The probability of a Z value greater than 1.5 would be .0668 (i.e., $1 - .9332$). To find the probability of a Z value being more than $+1.5$ or less than -1.5 (in other words, more than 1.5 standard deviations removed from the mean), we would merely double this figure and get .1336.

From Table B.2, we can verify that for a normal distribution there is a .6826 probability that an observation will fall within one standard deviation of the mean, a .9554 probability that it will be within two standard deviations, and a .9974 probability that it will be within three standard deviations.

An example may help clarify some of these ideas. ABC is a brokerage house that has a long-running program to train new brokers. In addition to interviews, the firm administers a test to decide which candidates will be accepted into the program. After testing thousands of candidates over the years they have found the scores are approximately normally distributed, with a mean of 70 and a standard deviation of 10. Given these facts, try the following questions:

1. What is the probability a new applicant taking the test will get a score above 92 (assuming we are not given any additional information about the person)?
2. What is the probability the applicant will get a score between 50 and 80?

Give it a try before reading on.

Answers

1. $$Z = \frac{X - \overline{X}}{\sigma}$$

$$Z = \frac{92 - 70}{10} = 2.2$$

Checking Table B.2, we see that the probability value corresponding to $Z = 2.2$ is .4861. Thus, there is a .9861 probability that a candidate will score 92 or less, or equivalently, a .0139 (1.39 percent) probability that the score will be higher.

TABLE B.2 Areas under the Normal Curve

An entry in the table is the proportion under the entire curve that is between *z* = 0 and a positive value of z. Areas for negative values of z are obtained by symmetry.

Second Decimal Place of *Z*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.0164	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Donald J. Koosis, *Business Statistics* (New York, NY: John Wiley & Sons, 1997). Copyright © 1997 by John Wiley & Sons; reprinted by permission.

2. This question is not as easy. It would be incorrect to proceed as follows:

$$Z = \frac{80 - 50}{10} = 3.0$$

Why? Because Z values must be measured relative to the mean. So the solution requires two steps: First, the probability of getting a score between 70 and 80 must be calculated. This can be done as follows:

$$Z = \frac{80 - 70}{10} = 1.0$$

Checking Table B.2, we find that this probability equals .3413. Next, to calculate the probability of a score between 50 and 70, we proceed as follows:

$$Z = \frac{50 - 70}{10} = -2.0$$

This corresponds to a probability of .4772. Thus, the probability of a score between 50 and 80 is the sum of these two values:

.3413 + .4772 = .8185 (81.85 percent)

Populations and Samples

If a data set contains all possible observations, it is called a *population*. If it consists of only a portion of these observations, it is called a *sample*. Whether a data set represents a population or a sample depends on the intended use. For example, if we are interested in the average income of all the employed people in Manhattan, the population would consist of all workers in Manhattan, and a sample would be only a portion of those workers. However, if we wish to estimate the average income of all U.S. workers, all workers in Manhattan would be a sample.

Intuitively, it should be clear that all workers in Manhattan would not be a very good sample of all U.S. workers. The problem in this case is that the sample is not representative of the population. In order for a sample to be representative of a population, it must be a *random sample*. A random sampling process is one in which each sample that can be drawn from the population has an equal chance of being selected. Samples that are not random will be *biased*, and a sampling approach that is not random will yield biased estimates. The mean of sample means that are biased will deviate from the population mean. Ironically, for a biased sample, the larger the sample size, the more certain its mean will deviate from the population mean.

In standard terminology, when a measure refers to the population, it is called a *parameter*.[6] A measure that refers to a sample is called a *statistic*. Thus, the standard deviation for a population (σ) is a parameter, and the standard deviation of a sample (s) is a statistic.

[6] The meaning of the term *parameter* when used in this context should not be confused with the distinction among *parameters*, *variables*, and *constants* explained in footnote 2 of Appendix A.

Estimating the Population Mean and Standard Deviation from the Sample Statistics

Although the intention of probability testing is to draw inferences about a population, it is usually impractical to collect data for the entire population. In fact, it is frequently impossible, since some populations are infinite. For example, the number of heads in 10 tosses of a coin is an infinite population, since there is no limit to how many times this event can be repeated. In practice, most applications of probability testing, including those in regression analysis, are based on samples rather than on populations.

Thus far, we have avoided the troublesome fact that the population mean and standard deviation are usually not known. We must now turn to the question of how the population mean and standard deviation can be estimated from a sample. It can be demonstrated that the mean of a random sample is an unbiased estimate of the population mean, even if the population does not show a normal distribution. This is equivalent to saying that, on average, the mean of randomly selected samples will equal the population mean. The sample standard deviation, however, is not an unbiased estimate of the population standard deviation, since it tends to slightly underestimate the population parameter. It has been proved that an unbiased estimate of the population variance (once again, variance is the square of the standard deviation) is given by the following equation[7]:

$$s^2 = \frac{\sum \left(X - \bar{X} \right)^2}{n - 1}$$

Taking the square root to translate this variance into a standard deviation, we get

$$s = \sqrt{\frac{\sum \left(X - \bar{X} \right)^2}{n - 1}}$$

This formula is almost identical to the population standard deviation. The only difference is the use of the divisor $n - 1$ instead of N.[8] For large samples, the difference between the formulas will be nearly negligible.

Finally, although the sample mean is an unbiased estimate of the population mean, this does not suggest the sample mean is necessarily close to the population mean. Thus, in addition to the point estimate provided by the sample mean, it would be highly desirable to determine a probable range for the population mean. But before we consider how such a range might be determined, we must first grasp the concept of a sampling distribution.

[7] When a standard deviation refers to a sample rather than a population, it is designated by an s instead of σ.

[8] The quantity $n - 1$ is called the number of degrees of freedom. We will define this term later.

Sampling Distribution

Fast Fred is a relatively active day trader. Being meticulous—but old-fashioned—at the end of every trading day he records the details of each of his trades in a notebook because he feels doing so helps him better absorb the lessons of his successes and failures in the markets. He eventually realizes that he should have kept his entries in an Excel spreadsheet so he could make calculations on his performance, but being a creature of habit, he continues to enter his trades in his notebook.

Fast Fred varies the number of contracts per trade based on the volatility of the market. He does all his trades using market orders. Recently, he has noticed that his average slippage per trade has increased significantly. (Slippage is the difference between the actual execution price and the market price at the time of trade entry.) Being concerned that his trading approach may no longer be viable, Fast Fred begins monitoring his slippage and notices that it is running around $75 per trade, which he believes is roughly $50 higher than it has averaged in the past. He reasons that if his average net profit (profit after gross commission and slippage) is not at least $60 per trade, it is probably not worthwhile continuing to trade. Unfortunately, he has never bothered to compile summary statistics from his many trades. The thought of going through all his trade records, which he estimates at more than 3,000 for the past year alone, seems worse than just taking his chances. Instead, he decides to draw a sample.

Knowing a little about statistics, Fred creates a random sample of 30 trade entries and calculates the average net profit per trade of this sample is $85 and the standard deviation of the sample is $100. He believes a 95 percent probability of an expected gain of at least $60 per trade is necessary to justify his continued trading activity. (An implicit assumption is that the past mean gain can be used as an estimate of his future expected gain per trade.) Given this information, is Fred's day trading method still viable? Unfortunately, we are not quite ready to answer this question without some additional theoretical background.

We will eventually return to Fred's dilemma, but first let us consider what might happen if Fred took another random sample of all his trades (including those selected for the first sample).[9] The mean net profit per trade of this sample would be different. If he repeated this process many times, Fred would generate list of different means, each corresponding to a different sample. However, it should be apparent these sample means would be much less spread out (i.e., have a smaller standard deviation) than the individual observations in a single sample. As will be detailed shortly, the standard deviation of observations within a sample and the standard deviation of sample means are related in a specific way.

In Figure B.4, hypothetical sample means for the net profit per trade are grouped by class (ranges of $10), with the *y* axis indicating the frequency of occurrences in each class. If the

[9] The assumption that trades that were picked for a prior sample can be picked again is important. Remember, the definition of a random sample is that each sample must have an equal chance of being selected. If the trade entries are not replaced, all possible samples that included any of the original trades will no longer be able to be picked—violating the random sample assumption. If the population is very large, the absence of replacement will not be significant, since combinations involving the selected sample will account for only a minute fraction of all possible combinations.

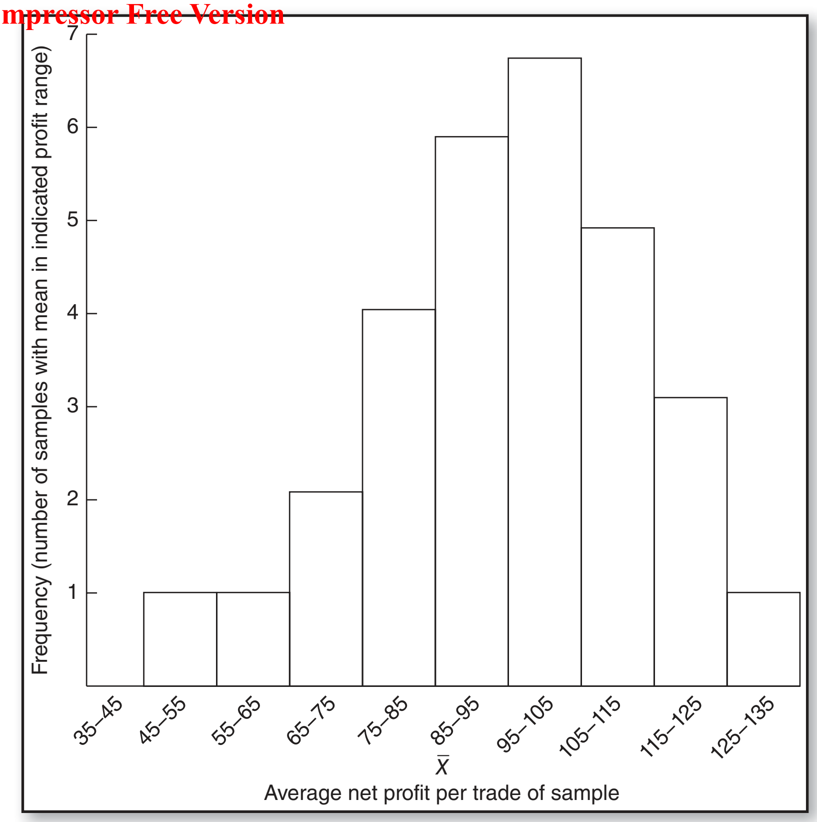


FIGURE B.4 Sampling Distribution of Mean

number of samples was repeated infinitely, and the class sizes were reduced correspondingly, Figure B.4 would approach a continuous curve known as a sampling distribution. The key point to realize is that the sampling distribution is a probability distribution curve related to sample statistics (e.g., sample means). Looking at Figure B.4, we might guess the sampling distribution would be similar to a normal distribution. In fact, if the sample size (i.e., standard size of each sample, not number of samples) is large enough, the sampling distribution will precisely approach a normal distribution.

Central Limit Theorem

The preceding illustration leads us to the *central limit theorem,* one of the most important concepts in statistical testing. The central limit theorem can be paraphrased as follows: The distribution of sample means from a population will approach a normal distribution as the sample size increases even if the population is not normally distributed.

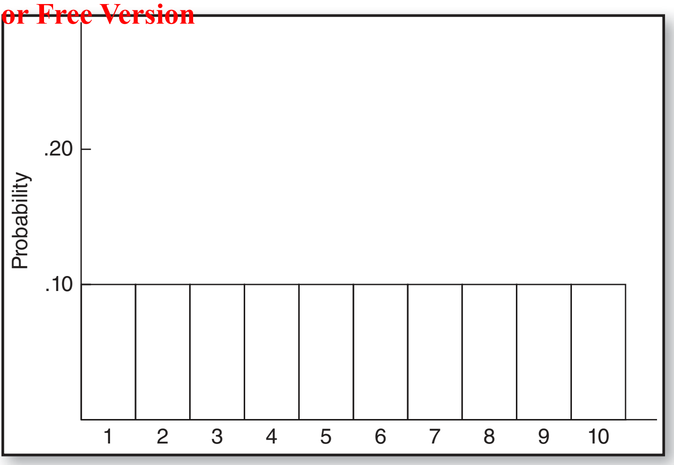


FIGURE B.5 Probability Distribution for Spinning Wheel

To illustrate the central limit theorem, consider the probability distribution for the number that turns up when spinning a wheel numbered 1 through 10. The probability distribution for this random variable is depicted in Figure B.5. Assuming an *honest* wheel, each number has an equal 0.10 probability of turning up. This illustration is obviously well removed from a normal probability

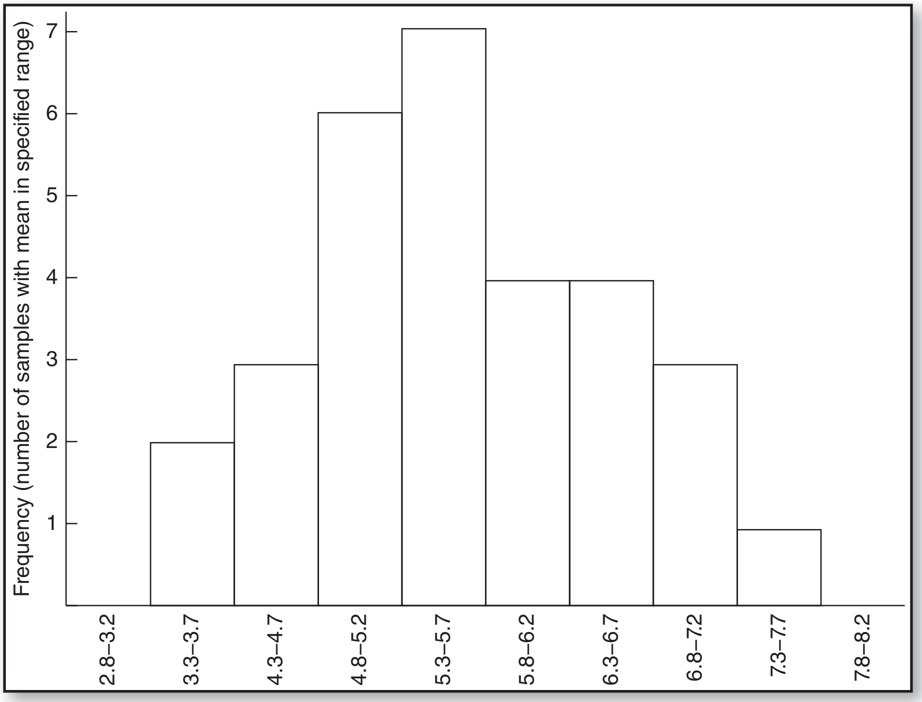


FIGURE B.6 Sampling Distribution of Mean for Spinning Wheel Trials

TABLE B.3 30 Samples on Spinning Wheel ($N = 10$)

Sample Number	Numbers on Wheel (10 spins)										Mean ($\overline{X}$)
1	8	10	5	6	6	2	4	6	8	10	6.5
2	5	7	1	1	4	3	8	9	5	3	4.5
3	8	5	4	10	7	5	5	4	10	10	6.8
4	3	1	8	5	7	1	6	5	9	10	5.5
5	1	9	10	9	3	2	6	5	2	10	5.7
6	9	1	6	2	1	3	5	7	3	1	3.8
7	4	6	6	10	8	4	4	9	5	2	5.8
8	4	10	10	2	4	5	6	3	8	1	5.3
9	8	7	8	10	6	6	10	3	1	9	6.8
10	7	4	9	8	6	9	7	6	8	10	7.4
11	7	9	2	10	3	7	10	5	10	9	7.2
12	6	4	1	3	8	8	1	1	10	7	4.9
13	5	7	2	7	9	6	4	8	8	9	6.5
14	1	2	6	10	3	5	10	9	1	4	5.1
15	7	4	10	6	8	2	4	5	4	3	5.3
16	5	3	1	10	3	10	7	4	7	5	5.5
17	6	2	4	8	8	5	8	5	4	8	5.8
18	6	3	9	2	4	9	9	6	1	10	5.9
19	2	5	3	6	9	3	4	6	6	9	5.3
20	6	2	1	8	6	1	5	2	9	7	4.7
21	4	4	5	7	8	7	5	10	8	6	6.4
22	2	9	10	6	9	1	4	5	3	5	5.4
23	5	4	7	1	10	1	4	7	3	3	4.5
24	9	4	5	2	6	9	6	4	2	2	4.9
25	4	5	8	5	7	6	8	5	9	7	6.4
26	8	2	1	2	8	6	8	7	1	6	4.9
27	7	8	7	6	6	5	1	7	9	6	6.2
28	9	7	7	5	9	4	3	3	2	1	4.9
29	2	3	5	7	9	1	6	1	8	9	5.1
30	4	3	2	9	2	1	8	4	1	6	4.0

distribution. Table B.3 summarizes the means of 30 samples of 10 spins each.[10] These samples are grouped by class in Figure B.6. Note the sample means roughly approximate a normal distribution, even though the parent population bears no resemblance to a normal distribution. Our sample size of 10 was fairly small. If a larger sample size had been used, the approximation of a normal distribution would have been better.

[10] These numbers were constructed using a random numbers table, an approach precisely equivalent to the example given.

Before moving on, bear in mind the examples involving repeated samplings were intended only as illustrations to elucidate the concepts of sampling distributions and the central limit theorem. In practice, however, we would always select only a single sample. Accuracy could be improved by simply increasing the size of this single sample.

Standard Error of the Mean

The standard deviation of sample means is usually smaller than the standard deviation of any given sample. The standard deviation of sample means is called the standard error of the mean and is represented by the symbol: $\sigma_{\overline{X}}$. (*Standard error* is a frequently used statistical term and can be interpreted as the standard deviation of the sampling distribution of the given statistic. In this case, the given statistic is the mean. Other types of standard error related to regression analysis are considered in Appendix C.) Given a distribution with a standard deviation σ, it can be proved that a random sample of size n has the following standard error of the mean:[11]

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Of course, we usually will not know the value of σ and will have to use s as an unbiased estimate of σ. (Recall that the two are very similar for all but very small samples.) Thus, in practice we would use

$$\sigma_{\overline{x}} = \frac{s}{\sqrt{n}}$$

For example, if the standard deviation of the sample (s) is 20 and the sample size is 25, then $\sigma_{\overline{x}}$ would equal 4. The larger the sample, the smaller $\sigma_{\overline{x}}$. However, note that the accuracy of the sample increases much more slowly than the sample size. For instance, a 25-fold increase in the sample size would reduce $\sigma_{\overline{x}}$ only by a factor of 5.

Confidence Intervals

Recall that assuming a data set is normally distributed, the probability of an observation falling within a given range can be determined from Table B.2. For example, the $\pm Z$ values that include 95 percent of observations are ± 1.96, since 2.5 percent of the distribution lies above $+1.96$ and 2.5 percent below -1.96. (Table B.2 indicates that .4750 of the area lies between $Z = 0$ and $Z = +1.96$; so, given

[11] This formula applies to infinite populations or samples in which the sample size is relatively small compared with the population. Although we will not be concerned with such cases, the precise formula when the sample size represents a significant percent of the population is

$$(\sigma / \sqrt{n})\sqrt{(N - n)/(N - 1)}$$

where n = sample size and N = population size.

the symmetry of the normal distribution, 95 percent of observations could be expected to fall within the range of −1.96 to +1.96.)

The formula for the Z value was formerly stated as

$$Z = \frac{X - \overline{X}}{\sigma}$$

In the case of a distribution of sample means (which the central limit theorem assures us will approximate a normal distribution), we have

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{x}}}$$

where $\overline{X}$ = sample mean

μ = population mean

$\sigma_{\overline{x}}$ = standard error of the mean (i.e., the standard deviation of sample means)

From the previous section, we know that $\sigma_{\overline{x}}$ can be approximated by $s/\sqrt{n}$. Thus,

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

or

$$\mu = \overline{X} - Z \cdot \frac{s}{\sqrt{n}}$$

If we were interested in the area that enclosed 95 percent of sample means, Z = ±1.96, the previous formula could be expressed as

$$\mu = \overline{X} \pm 1.96 \frac{s}{\sqrt{n}}$$

$$\overline{X} - 1.96 \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + 1.96 \cdot \frac{s}{\sqrt{n}}$$

This calculation can be interpreted as follows. In repeated samplings the true population mean could be expected to lie between $\overline{X} - 1.96 \cdot s/\sqrt{n}$ and $\overline{X} + 1.96 \cdot s/\sqrt{n}$ 95 percent of the time. Such a range is called a *confidence interval.*

The confidence interval can be used to test hypotheses about the population mean.[12] The standard approach involves testing the null hypothesis, which states there is no difference between the sample mean and the hypothesized population mean. Typically, we want to reject the null hypothesis, or, equivalently, demonstrate the sample mean is different from the hypothesized population mean at

[12] This discussion refers to population and sample means. However, it applies more generally to any sample statistic used to test an hypothesis about the population parameter.

some specified level of significance. The most commonly used level of significance is 0.05 (5 percent), which means that the sample mean lies outside the 95 percent confidence interval of the hypothesized population mean.[13] A statistical rejection of the null hypothesis demonstrates, with a probability at the stated level, that the sample could not have been drawn from a parent population with the hypothesized mean.

Sometimes, however, it is more critical to minimize the chance of rejecting the null hypothesis when in fact it is true (i.e., accepting that the sample mean is statistically different from the hypothesized population mean when it is not).[14] In such a case we might use a 0.01 level of significance. Of course, there is a tradeoff, because the lower the value for the level of significance (the more stringent the test), the wider (less specific) the confidence interval.

The *t*-Test

The *Z*-test is appropriate when the sampling distribution is normal, a condition that can be assumed true when the sample size is large.[15] However, for small samples the sampling distribution is better approximated by the *t*-distribution, and hence the *t*-test is more accurate. The *t* distribution is very similar to the normal distribution for all but very small samples. As the sample size increases, the normal and *t* distributions become increasingly similar. For example, at a 0.05 level of significance for a one-tailed test, the *t* value is 10 percent greater than the *Z* value for a sample of 10, 3 percent greater for a sample of 30, and 1 percent greater for a sample of 100. For an infinite sample, the normal and *t* distributions will be identical.

Similar to the standardized normal distribution, the *t* distribution is symmetrical, with a mean equal to zero and a standard deviation equal to 1. The formula for the *t* value of a sample statistic (e.g., mean) is totally analogous to the *Z* value:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

[13] This statement assumes that there is no a priori reason for assuming a value above or below the hypothesized mean. Such a situation is referred to as a two-tailed test. If, however, there is reason to believe that the sample mean would be above the null hypothesis population mean, the relevant question would be whether the sample mean was significantly *higher* than the population mean, not whether it was significantly *different from* the population mean. Such a situation is called a *one-tailed test*. The 0.05 significance level for a one-tailed test would correspond to the probability that a value was outside the 90 percent confidence interval. The distinction between one-tailed and two-tailed tests is discussed in greater detail in subsequent sections.

[14] An incorrect decision of this type is called a *type 1 error.* The probability of making a type 1 error is indicated by the level of significance. Accepting the null hypothesis when it is false is called a *type 2 error.* It should be stressed that the acceptance of the null hypothesis does not prove it is true, but only indicates that the null hypothesis could not be rejected at the stated level of significance. Thus, the acceptance of the null hypothesis does not prove that the sample was drawn from a population with the hypothesized mean, but rather that the sample and hypothesized population means are not statistically different at the specified level of significance.

[15] The meaning of *large* depends on the distribution of the underlying population. Roughly speaking, 30 is usually sufficiently large.

The *t*-test uses the *t* distribution for probability testing and is entirely analogous to the *Z*-test.[16]

The specific *t* distribution will depend on the *degrees of freedom* (*df*)—the number of observations (sample size) minus the number of constraints. For example, in tests of the sampling distribution of the mean, $df = n - 1$. There is one constraint, since given the mean, only $n - 1$ terms can be freely assigned. To see this, assume we have 10 observations with a mean of 50. If the sum of the first nine items equals 400, the value of the last term must be 100. Thus we say there are only $n - 1$ *df.* In a two-variable regression line, there are two parameters: *a* and *b.* Once these are fixed, only $n - 2$ terms can be assigned freely. Thus, *t*-tests of regression coefficients in the two-variable model are based on $n - 2$ degrees of freedom.

The application of the *t*-test is almost totally analogous to the *Z*-test. The only difference between the two is that the specific value used in the *t*-test depends on the degrees of freedom. Table B.4 provides a list of *t* values. The appropriate row is determined by the number of degrees of freedom, and the column by the desired level of significance in testing. Given the great similarity between the *Z*-test and the *t*-test, it would probably be redundant to provide a detailed description of the use of Table B.4. However, to check that you understand how to use this table, try the following questions:

1. If you are testing the hypothesis that the population mean is not significantly *greater than* the null hypothesis, what value must *t* exceed to reject this hypothesis at a 0.05 level of significance (i.e., to conclude that the true population mean is significantly greater than the null hypothesis)? The sample size is 20.
2. If you are testing the hypothesis that the population mean is not significantly *different from* the null hypothesis, what value must *t* exceed in order to reject this hypothesis at the 0.05 level of significance (i.e., to conclude that the true population mean is significantly different from the null hypothesis)? Once again, the sample size is 20.
3. a. Given a four-unit sample with a mean equal to 40 and a standard deviation equal to 10, what is the 95 percent confidence interval for the population mean?
   b. Now try it for a sample size equal to 30.

Answers

1. 1.729. For $df = 19$, Table B.4 indicates that there is only a 5 percent probability that this level will be exceeded. This type of test is called a one-tailed test.
2. 2.093. A 5 percent probability of being significantly different from the null hypothesis is equivalent to determining the *t* values that will define the boundaries for the upper and lower 2.5 percent of the distribution. This is an example of a two-tailed test.

[16] The astute reader may well wonder why we bother describing the *Z*-test in the first place, since the *t*-test would be more accurate for samples. The reason is that the mathematics underlying the *t* distribution assume that the population of the data series is normally distributed. This is a much stronger assumption than was necessary for the application of the *Z*-test, which only required that the sampling distribution be normal—a condition that the central limit theorem guaranteed would be approximately fulfilled for a sufficiently large sample. Thus, the *Z*-test provides the justification for probability testing of non-normally distributed populations. This is a critical fact, since the assumption of a normally distributed population is often not warranted.

TABLE B.4 **Student's *t* Distribution**

The first column lists the number of degrees of freedom (*k*). The headings of the other columns give probabilities (*P*) for *t* to exceed the entry value. Use symmetry for negative *t* values.

	P				
df	**.10**	**.05**	**.025**	**.01**	**.005**
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Source: Donald J. Koosis, *Business Statistics* (New York, NY: John Wiley & Sons, 1997). Copyright © 1997 by John Wiley & Sons; reprinted by permission.

3. a. $\bar{X} \cdot t \cdot \frac{s}{\sqrt{n}} < \mu < \bar{X} + t \cdot \frac{s}{\sqrt{n}}$

$$40 - 3.182 \cdot \frac{10}{\sqrt{4}} < \mu < 40 + 3.182 \cdot \frac{10}{\sqrt{4}}$$

$$24.09 < \mu < 55.91$$

b. $40 - 2.045 \cdot \frac{10}{\sqrt{30}} < \mu < 40 + 2.045 \cdot \frac{10}{\sqrt{30}}$

$$36.27 < \mu < 43.73$$

Note how dramatically the larger sample size increases the precision of the estimated confidence interval at the same probability level.

The choice of whether to employ a one-tailed or two-tailed test is not always clear-cut. Normally, a two-tailed test is used when we do not have any preconceived conclusion about the sample. In this case, the probability test for significance must allow for variation in either direction of the statistic being estimated (e.g., population mean). However, sometimes there are strong reasons to believe the sample statistic will be above or below the hypothesized population value—the only question being whether the difference will be significant. This type of situation will often apply in testing the significance of regression coefficients, as will be detailed in Appendix C.

It is finally time to return to our beleaguered day trader. We now see the solution to Fred's dilemma is fairly straightforward. You might wish to return to the section, "Sampling Distribution," to try to determine the correct decision before reading on.

Given the previously stated assumptions, the confidence interval for the expected net profit per trade would be

$$\$85 - 1.699 \cdot \frac{\$100}{\sqrt{30}} < \text{expected net profit per trade} < \$85 + 1.699 \cdot \frac{\$100}{\sqrt{30}}$$
$$\$53.98 < \text{expected net profit per trade} < \$116.02$$

Thus, it is not possible to say that there is a 95 percent probability the expected net profit per trade is greater than $60.

A few comments are in order. First, a one-tailed test is used because Fred is only concerned about testing the statistical significance of the expected net profit per trade being *greater than* $60 rather than the statistical significance of it being *different from* $60. Second, it should be stressed the confidence interval merely failed to demonstrate with a 95 percent or higher probability that the population expected net profit per trade was more than $60; it in no way proved that this figure was less than $60. Such a proof would have required a sample mean of $28.97 (or less), which would have implied a confidence interval of −$2.05 to $59.99. Third, if Fred had chosen a less-restrictive probability requirement, such as 90 percent, the confidence interval would have been

$$\$85 - 1.311 \cdot \frac{\$100}{\sqrt{30}} < \text{expected net profit per trade} < \$85 + 1.311 \cdot \frac{\$100}{\sqrt{30}}$$
$$\$61.06 < \text{expected net profit per trade} < \$108.94$$

implying the opposite decision.

The arbitrariness of the preceding example might seem unsettling. However, it should be emphasized the tester is free to choose the criterion that is deemed most important. If Fred is very concerned about continuing to trade when in fact such a decision would be unwarranted by the true population expected net profit per trade (type 1 error), he would choose a low (restrictive) value for the level of significance for testing. If he was less concerned about this type of error, he would use a higher value for the level of significance. In fact, if Fred's primary concern was to avoid terminating his trading when the true expected net profit per trade was actually more than $60, he might continue to trade even if the sample mean was less than $60, constructing a confidence interval to test whether the sample mean was significantly below the hypothesized $60 population mean.

Checking the Significance of the Regression Equation

Factual evidence can never "prove" a hypothesis; it can only fail to disprove it, which is what we generally mean when we say, somewhat inexactly, that the hypothesis is "confirmed" by experience.

—Milton Friedman

The Population Regression Line

In Appendix A we discussed the derivation of a regression line on the basis of empirical data. While it was premature to raise the subject then, the fitted line provided by the regression formula is actually a sample of the true population line, which is not known. For example, the regression line relating hog slaughter to the pig crop was a sample of the true relationship between the two variables. The fitted line is a sample because it represents only one realization of an entire series of possible regression lines. The actual regression line realized will depend on measurement error in the data and the unknown influence of variables not included in the model.

The population or true regression model can be expressed as

$$Y_i = a + \beta X_i + e$$

where e is a randomly distributed *error* or *disturbance term.*

Even if we knew the true population regression line, the actual observed values Y_i would still deviate from the predicted level by an amount equal to the error term *e*. The key reason for this is that a regression equation is a highly simplified model for the behavior of the dependent variable. In reality, the number of hogs slaughtered will depend on many more variables than just the pig crop—for example, the distribution of pigs born during the period, weather conditions, feed prices, and hog prices. Although the magnitude of the disturbance term can be reduced by including other relevant variables in the regression model (this anticipates multiple regression, discussed in Appendix D), it is impossible to introduce enough variables to eliminate these deviations completely.[1]

In addition, even if all relevant variables were included in the model, observations would still deviate from the regression line due to measurement errors. This is not a trivial consideration, since data items that can be precisely measured (e.g., temperature) are by far the exception. Most data can only be estimated from samples (e.g., pig crop, hog slaughter).

Basic Assumptions of Regression Analysis

Appendix A explained that a basic assumption of regression analysis is that the relationship between the dependent variable *Y* and the independent variable *X* is linear. Several other key assumptions are related to the error terms:

1. The mean value of the error terms equals zero.
2. The error terms have a constant variance equal to σ^2.
3. The error terms are independent random variables. This assumption has two important implications:
    a. The error terms are uncorrelated.
    b. The error terms and the independent variable *X* are uncorrelated.
4. The error terms are normally distributed.

These assumptions underlie the various tests used to assess a regression model's reliability.

Testing the Significance of the Regression Coefficients

The empirically derived values of *a* and *b* will not equal the population values of α and β except by chance (Figure C.1). It can be shown, however, that *a* is an unbiased estimate of α and *b* is an unbiased estimator of β.[2] Actually, it can be demonstrated that *a* and *b* are not only unbiased estimators, but

[1] Even if all such variables were known and could be precisely determined—two extraordinarily unlikely assumptions—the regression computation would still limit the number of variables that could be introduced, since each additional variable would reduce the degrees of freedom by 1. The significance of the regression equation is reduced as the degrees of freedom decline, and the equation becomes totally trivial when the number of variables is equal to the number of observations.

[2] An unbiased estimator is one that on the average will equal the population parameter. In other words, the mean of the sampling distribution for an unbiased estimator will be equal to the population parameter value.

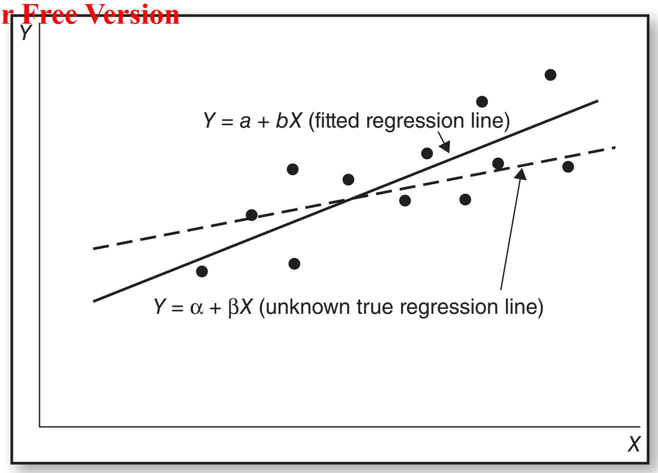


FIGURE C.1 Fitted vs. True Regression Line

that they are the "best linear unbiased estimators" (BLUE). This means that *a* and *b* will have the lowest variance (i.e., are the most "efficient") among all possible unbiased *linear* estimators of α and β.

Although *b* provides an unbiased point estimate for the population regression coefficient β, we would like to know the variability of this estimate. In other words, we are interested in the standard error of *b*. Recall from Appendix B that the standard error is the standard deviation of a sampling distribution of a statistic. In this case, the relevant statistic is the regression coefficient *b*.

A diagram can help clarify this point. Figure C.2 shows a distribution with a mean equal to β—the population regression coefficient. Figure C.2 illustrates the distribution that will be formed by the estimated values of *b* if an infinite number of samples were drawn. In other words, Figure C.2 is a sampling distribution of *b*. The standard deviation of this distribution is called the standard error of *b*.

In Appendix B we indicated that the *t* value for a sample mean could be expressed as

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

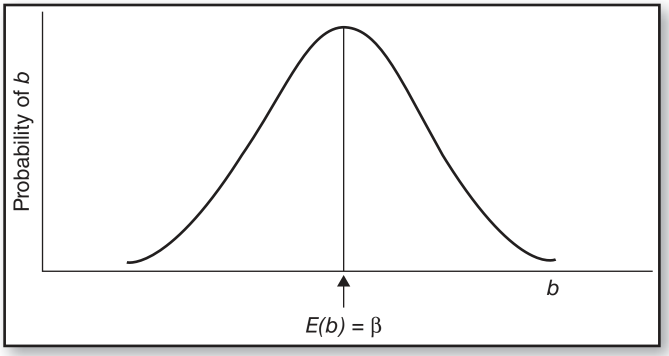


FIGURE C.2 Sampling Distribution of Regression Coefficient

The only difference now is that we are trying to make judgments about the population regression coefficient β from the sample coefficient b, rather than making decisions about the population mean μ from the sample mean $\overline{X}$. In general terms, the t value could be expressed as

$$t = \frac{\text{sample statistic} - \text{population parameter}}{\text{standard error (s.e.) of sample statistic}}$$

In other words, in all applications, the t values indicate the number of standard deviations between the sample statistic and the hypothesized population parameter. A large t value (approximately 2 or more) will indicate there is only a small probability the hypothesized population parameter could be correct. The higher the t value, the less likely the sample came from a parent population with the hypothesized parameter value. In the case of the regression coefficient, the preceding general formula for the t value could be expressed as follows[3]:

$$t = \frac{b - \beta}{\text{s.e.}(b)}$$

Frequently, we will be interested in testing the hypothesis that $\beta = 0$. The reason? In the absence of any information, the best estimate for a variable is the mean, or equivalently, the regression line

$$Y = a + bX$$

where $a = \overline{Y}$
$b = 0$

If, however, the independent variable has some explanatory power, then a regression line with a nonzero slope would offer a better fit to the observations (Y_i). Thus, a key question to be considered in any regression analysis is whether the regression coefficient b is significantly different from zero.

In order to answer this question, we test the assumption that the population regression coefficient $\beta = 0$. In this case the t value reduces to

$$t = \frac{b}{\text{s.e.}(b)}$$

Thus, the t value is the regression coefficient divided by the standard error of the regression coefficient. In other words, the t value indicates how many standard deviations the regression coefficient is from the population coefficient if our hypothesis that $\beta = 0$ were true. If the t value is high (e.g., an

[3] Strictly speaking, the distribution of b is not generally normal except in the limit of a large number of observations. Although this implies that the use of the t distribution for calculating confidence intervals is imprecise, from a practical standpoint, the t distribution will yield satisfactory results because the exact boundaries of the confidence interval are not critically dependent on the actual distribution of b.

absolute value approximately greater than 2.0),[4] it suggests that the population regression coefficient is not equal to zero.

In order to apply the preceding formula, we need to know the value of s.e.(b). Given the previously detailed assumptions of regression analysis, it can be demonstrated that

$$\text{s.e.}(b) = \sqrt{\frac{s^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

where s^2 is an unbiased estimator of the population variance σ^2. (σ^2 is the variance of the error terms, or equivalently, the variance of the observations from the unknown population regression line.[5]) Since the true regression line is not known, σ^2 must be estimated. It can be proved that s^2 is an unbiased estimator of σ^2 where[6]

$$s^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n}\left(Y_i - a + bX_i\right)^2}{n-2}$$

where Y_i = the individual observations

$\hat{Y}_i$ = fitted values (i.e., values implied by regression line) at X_i (the values of the independent variable corresponding to the observations Y_i)

Assuming that we are testing the hypothesis $b = 0$, we can now express the regression coefficient t value:

$$t = \frac{b}{\text{s.e.}(b)} = \frac{b}{\dfrac{s}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}} = (b)\frac{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}}$$

[4] For example, if there are 5 degrees of freedom, a t value of 2.0 would imply that a coefficient at least as much greater than 0 than the one measured would occur only 5 percent of the time, if the true population regression coefficient were equal to 0 (Table B.4). If there are 60 degrees of freedom, such an event would occur only 2.5 percent of the time. These figures assume that a one-tailed test is employed. (See discussion at end of this section.)

[5] Note that error terms refer to the differences between the observed values (Y_i) and the true population regression line (not the fitted regression line). *Terminology note:* The term *error* or *disturbance* describes the difference between an observation and the true population regression line (which is usually not known), while the term *residual* or *deviation* refers to the difference between an observation and the fitted regression line. This theoretical distinction is obscured by the rather commonplace use of error terms to refer to the residuals (the deviations between the observations and the fitted regression line).

[6] The reason for dividing by $n - 2$ is that 2 degrees of freedom are lost because of the constraints imposed by fitting a and b. In other words, for any given set of values for a and b, once $n - 2$ observations are specified, the remaining 2 observations can no longer be freely assigned. In general, the number of degrees of freedom will equal the number of observations minus the total number of parameters (see footnote 2 in Appendix A).

Note the following three facts:

1. The sign of the t value will depend upon the regression coefficient. If X and Y are inversely correlated, the regression coefficient and t will be negative. The sign of the t value is insignificant. We are only concerned with the absolute value of t in testing the significance of the regression coefficient b.
2. As seems intuitively desirable, the s term in the previous equation will ensure that the t value will decline as the sum of the squared deviations increases.
3. The narrower the range of X values for the observations, the lower the t value, since

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

will be smaller, hence the less reliable the regression coefficient estimate. This concept is illustrated in Figure C.3. Note that when the observations correspond to a small range of the independent variable (Figure C.3a), the influence of the deviations can easily swamp the effect of the slope, and the estimated regression line will not be very reliable. Conversely, when the observations correspond to a wide range of X values (Figure C.3b) the estimated regression line will be far more reliable.

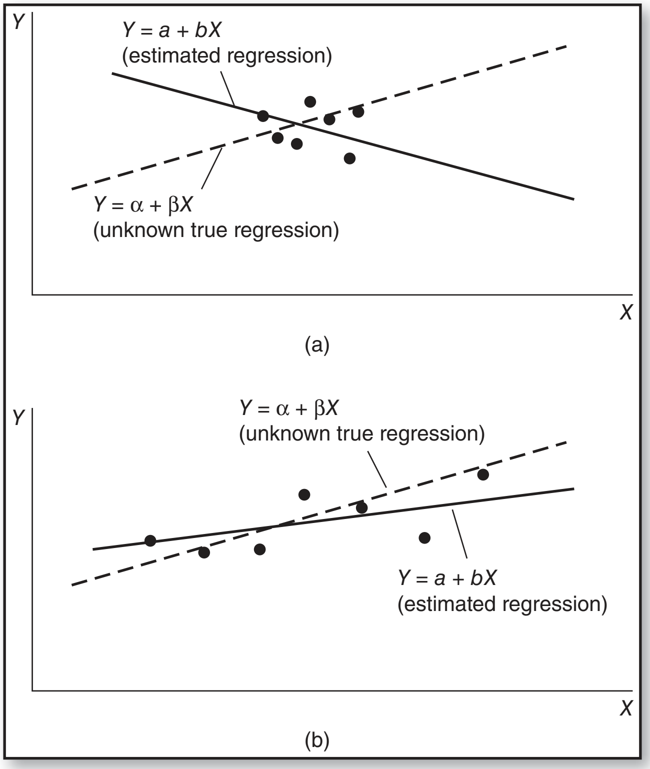


FIGURE C.3 Effect of Range of X_t on Reliability of the Regression Coefficient
Source: T. H. Wonnacott and R. J. Wonnacott. *Econometrics,* John Wiley & Sons, New York, 1980. Copyright © 1980 by John Wiley & Sons; reprinted by permission.

In testing the significance of the regression coefficient b, it is often more appropriate to use a one-tailed test. The reason for this is that the direction of the relationship between the dependent and explanatory variable(s) is usually known a priori. In such cases it is more relevant to test whether the coefficient value is significantly *greater than* or *less than* zero rather than whether it is significantly *different from* zero. For example, we know that a larger pig crop will result in higher hog slaughter. The only relevant question is if the relationship is statistically significant. Thus, the appropriate question is not whether b is significantly different from zero, but whether it is significantly greater than zero, since we do not entertain the possibility that a larger pig crop will result in lower hog slaughter. If, however, we wanted to test whether there was a relationship between the dependent variable and the independent variable, without any bias about the direction of the relationship, a two-tailed test would be used.

The t-test can also be applied to the constant or intercept term, a. In this case, the t value to test the hypothesis that the population regression line intercept is equal to zero would be

$$t = \frac{a}{\text{s.e.}(a)} = \frac{a}{s\sqrt{\dfrac{1}{n} + \dfrac{\bar{X}^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}}$$

In practice, there is usually little reason to be overly concerned about the significance of the constant term, and t-tests of the constant term can be omitted without any great loss.

As an example of how the t-test might be applied, consider the regression equation

$$Y = 1.1094\text{X} - 6.8276$$

derived in Appendix A. Table C.1 illustrates how to compute the t value for the regression coefficient b. Although it will not be necessary to compute the t value in practice, since the availability of a standard computer regression program is presumed, it is important to have a feel for the computations that underlie the key regression statistics. Checking the t value derived in Table C.1, we find that it far exceeds 1.734—the t value at a 0.05 level of significance for a one-tailed test with 18*df* (see Table B.4). We conclude the December–May pig crop is indeed significant in explaining the June–November slaughter.

The conclusion that the regression coefficient in the previous example is statistically significant might seem somewhat trivial. After all, the same conclusion would seem to be intuitively obvious by examining the scatter chart for the observations (see Figure A.4). In fact, as a generalization, unless one demonstrates an extraordinarily poor sense of intuition in choosing the independent variable in a *simple regression*, that is, a regression equation with only one explanatory variable, the t-test will usually prove significant. However, the t-test becomes critically important in evaluating a *multiple regression model*—a regression equation with two or more explanatory variables. In this case, a simple graphic depiction of the regression fit is no longer possible, and the significance of additional variables is often not intuitively apparent. The t-test is one of the most important statistical tests in regression analysis and will be considered further in the multiple regression case.

TABLE C.1 (*a*) Computing the *t*-Value for the Regression Coefficient

Year	Jun–Nov Hog Slaughter Y_i	Dec–May Pig Crop X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	Fitted Value $\hat{Y}_i$	Residual $Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$Y_i - \bar{Y}$	$(Y_i - \hat{Y})$
1995	50.077	48.2936	–2.726	7.429	49.529	–1.235	1.526	–4.128	17.037
1996	47.888	45.4527	–4.915	24.154	47.205	–1.753	3.071	–6.968	48.559
1997	48.394	46.2009	–4.409	19.437	47.742	–1.541	2.376	–6.220	38.692
1998	52.469	50.9291	–0.334	0.111	52.068	–1.139	1.297	–1.492	2.226
1999	51.519	51.1112	–1.284	1.648	51.060	0.052	0.003	–1.310	1.716
2000	50.087	49.6888	–2.716	7.375	49.539	0.149	0.022	–2.732	7.466
2001	49.472	49.1693	–3.331	11.094	48.887	0.283	0.080	–3.252	10.575
2002	50.858	50.7086	–1.945	3.782	50.358	0.351	0.123	–1.713	2.933
2003	50.029	50.7581	–2.774	7.693	49.478	1.280	1.639	–1.663	2.766
2004	50.737	52.2648	–2.066	4.267	50.229	2.035	4.143	–0.156	0.024
2005	51.33	52.3326	–1.473	2.169	50.859	1.474	2.172	–0.089	0.008
2006	52.242	53.1504	–0.561	0.314	51.827	1.323	1.751	0.729	0.532
2007	54.266	55.5693	1.463	2.141	53.975	1.594	2.540	3.148	9.911
2008	57.019	57.6483	4.216	17.777	56.898	0.751	0.563	5.227	27.323
2009	57.564	57.3914	4.761	22.670	57.476	–0.085	0.007	4.970	24.703
2010	56.326	55.6805	3.523	12.414	56.162	–0.482	0.232	3.259	10.623
2011	57.118	56.2641	4.315	18.622	57.003	–0.739	0.546	3.843	14.768
2012	57.818	57.4775	5.015	25.153	57.746	–0.268	0.072	5.056	25.567
2013	57.02	55.9142	4.217	17.786	56.899	–0.985	0.969	3.493	12.201
2014	53.821	52.4176	1.018	1.037	53.503	–1.085	1.178	–0.004	0.000
	$\Sigma Y_i = 1{,}056.05$	$\Sigma X_i = 1{,}048.42$		$\Sigma\,(X_i - X)^2 = 207.073$			$\Sigma\,(Y_i - Y_i)^2 = 24.311$		$\Sigma\,(Y_i - Y)^2 = 257.630$
	$Y = 52.803$	$X = 52.421$							

Fitted regression line (from Table A.1): Y = –3.6279 + 1.0615X

$$t = \frac{b}{\text{s.e.}(b)} = \frac{b \cdot \sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}{\sqrt{\dfrac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}} = \frac{1.0615\sqrt{207.073}}{\sqrt{\dfrac{24.311}{18}}} = 13.144$$

(*b*) Computing r^2

Note: Ignore this section of the table until reaching the appropriate section later in this appendix.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{24.311}{257.63} = 0.9056$$

Standard Error of the Regression

The *standard error of the regression* (SER) is the standard deviation of the residuals, or equivalently, the standard deviation of the observations from the fitted regression line:[7]

$$\text{SER} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

This formula should look familiar. The SER was previously denoted by an s and appeared in the calculation of the standard error of the regression coefficient, s.e.(b). The name *standard error of the regression* merely highlights the fact that this figure is a measure of dispersion for the entire equation. Note that the wider the scatter of points from the regression line, the larger the SER.

It should be emphasized that the SER can only be interpreted relative to the range of the dependent variable. For example, in price-forecasting equations, an SER figure of 10¢ would be indicative of an excellent fit if the given price series ranged between \$6 and \$12, and an extremely poor fit if the price range were \$0.30 to \$0.60. For this reason, as long as the mean of the dependent variable $\overline{Y}$ is greater than the range between the high and low values, it may be useful to consider the percent (%SER) where[8]

$$\%\text{SER} = \frac{\text{SER}}{\overline{Y}}$$

In effect, the %SER is a dispersion measure normalized by the magnitude of the underlying data. When comparing different equations with the *same dependent variable*, over the same time interval, the %SER will yield the same conclusions as the SER, with the advantage of being stated in terms that are intuitively more meaningful.[9]

Confidence Interval for an Individual Forecast

Assume that we used our regression equation to forecast a future value of the dependent variable Y, given a specific value of the independent variable X. What forecast range would have a 95 percent probability of containing the true value of Y? (*Implicit assumption:* The estimated value for the

[7] The SER is also often referred to as the standard error of the estimate (SEE), or simply as the standard error (SE).

[8] If $\overline{Y}$ is less than the range, the %SER could be very misleading. For example, if the dependent variable included values that were both positive and negative, its mean could be close to 0. In this case, the %SER could approach infinity.

[9] Caution should be exercised in drawing any conclusions from comparisons that involve equations with different dependent variables, since the %SER would be sensitive to the dependent variable chosen. For an example of why this is undesirable, see the discussion in the section "Coefficient of Determination (r^2)."

independent variable is either known or precisely projected; that is, there is no forecast error involved in the X value.) In answering this question, we note that even if all the regression assumptions are fulfilled, there are three sources of potential error:

1. **Error of the mean.** The true population regression line is unknown and is estimated from the observations. The resulting fitted line passes through $(\overline{X}, \overline{Y})$, while the population line passes through $(\overline{X}, \overline{Y}_{\overline{X}})$, where in general, $\overline{Y}_{\overline{X}}$, the unknown population mean at $\overline{X}$, will not equal the sample mean $\overline{Y}$. As indicated in Figure C.4a, this type of error will result in all forecasts (i.e., projections of Y for any value of X) being too high or low by the difference between the value of $\overline{Y}$ and $\overline{Y}_{\overline{X}}$. (Figure C.4a depicts the case of a positive error of the mean; a symmetric image would apply for a negative error.)
2. **Error of the slope.** There will also be some difference between the true regression coefficient β and the fitted line slope b. Figure C.4b illustrates the combined effect of this source of error and the error of the mean for the forecasted value $\hat{Y}_f$ at X_f. Note that the error of the slope will be zero at $\overline{X}$ since the fitted regression must pass through $(\overline{X}, \overline{Y})$, but will increase steadily as X values are further removed from $\overline{X}$. (The illustration in Figure C.4b depicts the case of a positive error of the mean and an estimated regression coefficient that is too high; a symmetric image would apply for the reverse case.)
3. **Random error.** For reasons previously explained, even if the true population line were precisely known, there would still be error terms. The confidence interval for the population line is illustrated in Figure C.4c. Note that the interval is independent of the value of X.

The forecast error for any individual prediction would reflect the combined effect of all three of the influences just discussed (Figure C.4d). The shape of the confidence interval depicted in Figure C.4d is a consequence of the fact that, although the error of the mean and the random error components will be equal for all values of X, the error of the slope increases for values further removed from $\overline{X}$. If the independent variable value for the forecast period is denoted by X_f, the standard error for the forecasted value of Y_f is given by

$$\text{s.e.}(\hat{Y}_f) = s\sqrt{1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

Note that when X_f, equals $\overline{X}$, and n is large, this term reduces to approximately s—the standard deviation of the residuals. (Once again, s is more commonly called the standard error of the regression.) Also note that the further X_f is from $\overline{X}$, the larger the s.e. $(\hat{Y}_f)$

The confidence interval for Y_f would be

$$\hat{Y}_f - t\sum \text{s.e.}(\hat{Y}_f) < Y_f < \hat{Y}_f + t \cdot \text{s.e.}(\hat{Y}_f)$$

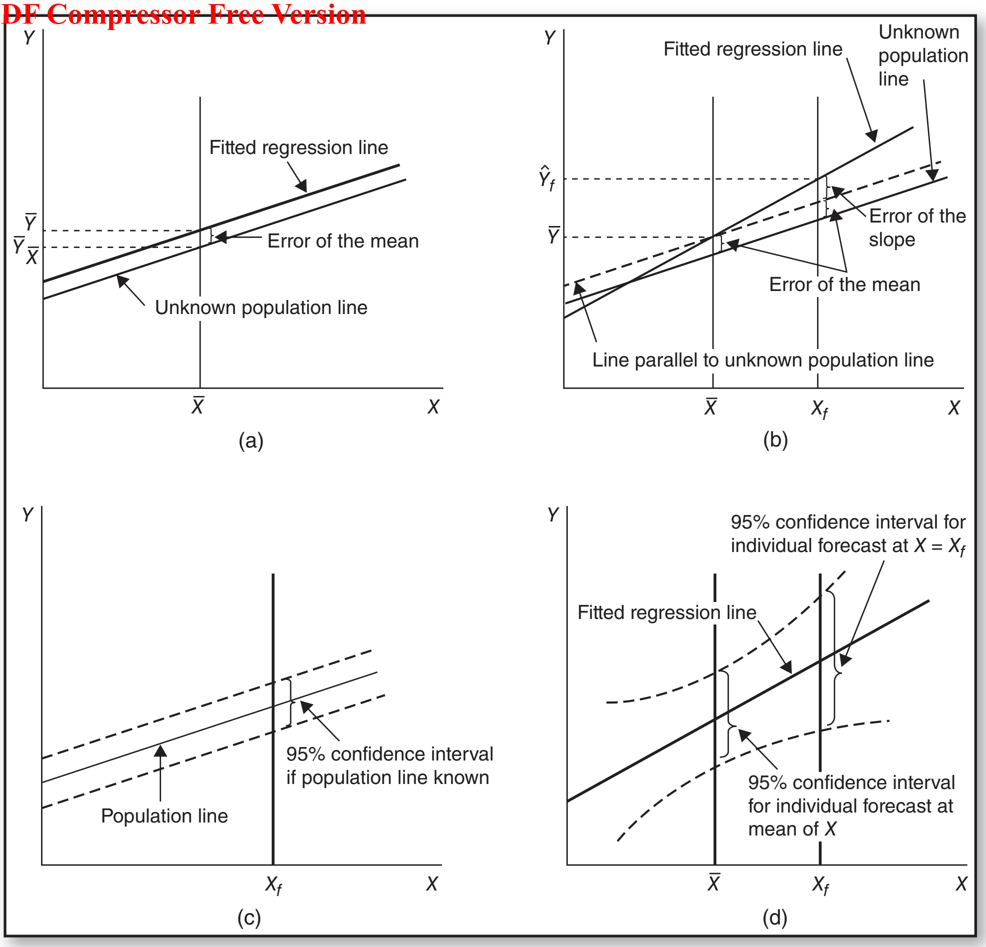


FIGURE C.4 Confidence Interval for an Individual Forecast

where $t = t$ value at the specified level of significance for the given number of degrees of freedom. At $X = \bar{X}$, this interval would reduce to

$$\hat{Y}_f - t \cdot s\sqrt{1 + \frac{1}{n}} < Y_f < \hat{Y}_f + t \cdot s\sqrt{1 + \frac{1}{n}}$$

Extrapolation

Extrapolation refers to predictions beyond the range of observations, that is, forecasts for Y_f where X_f is greater than or less than any observed value X_i. As should be apparent from the preceding formula for s.e. (Y_f), predictions in the extrapolated range will be particularly subject to uncertainty, since s.e. ($\hat{Y}_f$) increases as the X value moves farther away from $\overline{X}$. However, there is an even more important reason why extrapolation-based forecasts should be viewed with great skepticism. Basically, it is never safe to assume that the relationship between the dependent and independent variables exhibited in the observed range would continue to hold in the extrapolated range. For example, consider a market in which there is a rough inverse linear relationship between price and the final stock/consumption ratio during the observation period. This relationship would likely break down in a record shortage situation, since prices frequently begin to rise at an accelerated rate once an expected shortage reaches a critically low level.

What should be done in the case in which the expected value for the explanatory variable falls beyond the observed range? Many professional analysts faced with such a dilemma will apprehensively extrapolate and hope for the best (probably because of implicit pressure to provide forecasts, whether the necessary data exists or not). Such hopes are often misplaced. This does not mean to imply that the analyst must resign herself to a yearlong hibernation before she ventures any market forecast—few firms are that enlightened. Rather, the intended point is that in such a situation, the analyst must rely almost totally on her intuitive fundamental sense of the market and technical analysis to generate forecasts, rather than naively continuing to use a formal fundamental model that is no longer relevant.

Coefficient of Determination (r^2)

If the regression model were unavailable or, equivalently, if the independent variable were useless in explaining the dependent variable Y, the best forecast for a value of Y would be its mean $\overline{Y}$. We define the difference between an individual observation and the mean, $Y_i - \overline{Y}$, as the total deviation. Now if X is of any use in explaining Y, the deviations between the observed points and the fitted regression line should tend to be smaller than the total deviation. For any given observation Y_i, the portion of the total deviation explained by the regression equation would equal the fitted value minus the mean, $\hat{Y}_i - \overline{Y}$. The portion that remains unexplained will equal the observed value minus the fitted value $Y_i - \hat{Y}_i$. The relationship among the *explained*, *unexplained*, and *total deviations* is illustrated in Figure C.5. For any given observation, this relationship can be expressed as follows:

$$\begin{array}{ccccc} \text{Total deviation} & = & \text{explained deviation} & + & \text{unexplained deviation} \\ \text{for } Y_i & & \text{for } Y_i & & \text{for } Y_i \\ \left(Y_i - \overline{Y}\right) & = & \left(\hat{Y}_i - \overline{Y}\right) & + & \left(Y_i - \hat{Y}_i\right) \end{array}$$

If we are interested in deriving a relationship between *total variation*, a measure of the deviation of all points, and *explained* and *unexplained variation*, we cannot simply sum the terms, because opposite

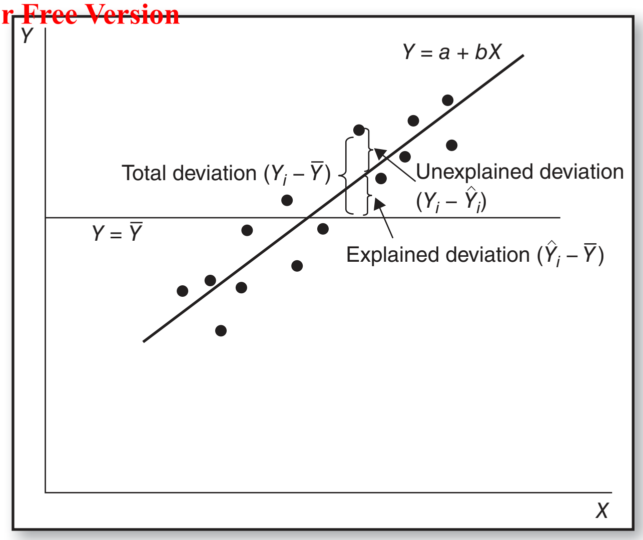


FIGURE C.5 Explained, Unexplained, and Total Deviation

deviations will offset each other, yielding a tautological relationship.[10] Thus, we square both sides of the equation before summing. This step is analogous to the approach used to get around the same problem in trying to find the best-fit line:

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 &= \sum_{i=1}^{n}\left[\left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)\right]^2 \\ &= \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 - 2\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right)\end{aligned}$$

Given the previously stated assumption that the independent variable X and the error terms are uncorrelated, it can be algebraically demonstrated that

$$2\sum_{i=1}^{n}\left(Y_i - \bar{Y}_i\right)\left(Y_i - \bar{Y}\right) = 0$$

[10] $$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) &= \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right) + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) \\ &= \sum_{i=1}^{n}\hat{Y}_i - \sum_{i=1}^{n}\bar{Y} + \sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}\hat{Y}_i \\ &= \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n}Y_i - n\bar{Y} \\ 0 &= 0\end{aligned}$$

Thus, we have

$$\text{Total variation} = \text{explained variation} + \text{unexplained variation}$$

$$\sum_{i=1}^{n}\left(Y_i - Y\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

where variation is defined as the sum of the squared deviations.

Dividing both sides by $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$

$$1 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} + \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

$$1 = \frac{\text{explained variation}}{\text{total variation}} + \frac{\text{unexplained variation}}{\text{total variation}}$$

We define r^2 as the

$$\frac{\text{Explained variation}}{\text{Total variation}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

or equivalently,

$$r^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

The second form is more convenient because the analysis of the regression equation focuses on the unexplained variation (residuals). Note that $0 \leq r^2 \leq 1$. If X does not explain any of the variation in Y, then $r^2 = 0$. If X explains all of the variation in Y (i.e., all the observations fall precisely on the fitted line), $r^2 = 1$. The r^2 calculation is an extremely useful summary statistic because of the significance of what it measures—the percentage of variation explained by the regression equation—and the intuitive clarity with which the r^2 figure can be interpreted. Table C.1*b* illustrates the calculation of r^2 for the regression line derived in Table A.1.

The statistic r^2 is extremely useful in comparing alternative models, as long as they all have the same dependent variable, that is, they differ only in the explanatory variables. However, when regression equations have different dependent variables, inferences based on r^2 can prove very misleading. For example, consider the following two models:

$$\text{Model I:} \quad P_t = a + bP_{t-1}, \quad r^2 = 0.98$$
$$\text{Model II:} \quad \Delta P_t = a + bX, \quad r^2 = 0.50$$

where P_t = closing price on day t
P_{t-1} = closing price on day $t - 1$
$\Delta P_t = P_t - P_{t-1}$
X = an explanatory variable known on day $t - 1$ and used to predict price changes

Despite the fact that Model I has a much higher r^2, Model II represents the better forecasting equation. If the survey period is sufficiently long, an equation such as Model I will merely tell us that the price on a given day will be almost equal to the preceding day's price. (In such an equation, b is likely to be very close to 1.0.) The reason for the high r^2 value is that although prices for the entire period may range widely, prices on adjacent days must be closely correlated (at most, they can be separated by the daily price limit). Model I, however, is totally useless in forecasting the next day's price. In contrast, the independent variable in Model II explains 50 percent of the price *change* on a given day and might be an extremely important aid in day trading. This illustration is intended to highlight the potential folly of making value judgments about regression equations with different dependent variables.[11]

[11] An even more extreme example is possible. If we used hogs not slaughtered (HNS) instead of hog slaughter (HS) as the dependent variable regressed against the pig crop (PC), where HNS = PC − HS, the sum of the squared residuals and hence SER would be exactly the same, but r^2 would be different. Why? Because r^2 would be affected by the dependent variable chosen

$$r^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

In this example,

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

will be the same whether HS or HNS is the dependent variable, but

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

will be different, hence r^2 will be different.

Spurious ("Nonsense") Correlations

It is important to understand that cause and effect in the regression procedure are in the eyes of the beholder. The r^2 value derived in Table C.1*b* merely tells us that there is a strong correlation between hog slaughter and the preceding pig crop. The way we interpret the cause and effect of this statistic emanates only from our theoretical understanding of the underlying process. In this particular example, it is quite obvious the pig crop in one period affects slaughter in the next period rather than vice versa. However, if enshrouded in ignorance, we set out to prove the level of hog slaughter determines the number of pigs born in the preceding period, the resulting equation would yield the identical r^2 value. Thus, r^2 only reflects the degree of correlation between two variables; it in no way proves a cause-and-effect relationship.

The potential folly of drawing cause-and-effect inferences from an r^2 figure is demonstrated by Figure C.6. Note what appears to be a striking relationship between the number of hedge funds and U.S. wine consumption. In fact, the r^2 value between the number of hedge funds and U.S. wine consumption during the period depicted is a remarkably high 0.99! What conclusions are we to draw from this chart?

- Increased wine consumption encourages people to invest in hedge funds.
- Hedge funds drive people to drink.
- The hedge fund industry should promote wine consumption.

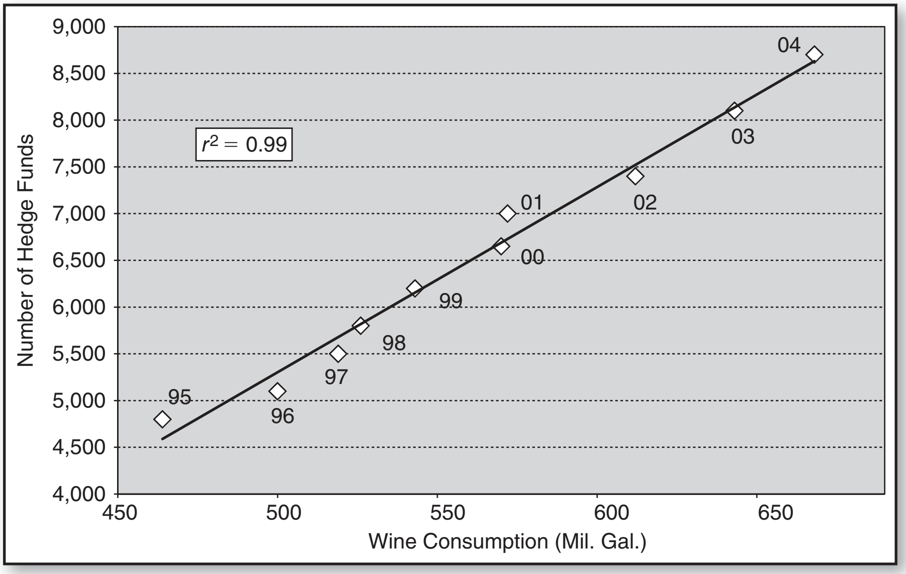


FIGURE C.6 Number of Hedge Funds vs. U.S. Wine Consumption

- Wine growers should promote hedge fund investing.
- All of the above.
- None of the above.

Actually, the striking correlation between wine consumption and the number of hedge funds is very easily explained. Both variables were affected by a common third variable during the period depicted: time. In other words, both the number of hedge funds and wine consumption witnessed pronounced growth trends during this time period. The apparent relationship arises from the fact that these trends were simultaneous. This type of coincident linear relationship is called "spurious" or "nonsense" correlation. Actually, the correlation is real enough; only the interpretation of cause and effect is nonsense.

The Multiple Regression Model

In our description of nature the purpose is not to disclose the real essence of the phenomena but only to track down, so far as it is possible, relations between the manifold aspects of our experience.

—Niels Bohr

Basics of Multiple Regression

In practice, it is rarely possible to explain the behavior of a dependent variable adequately with only one explanatory variable. For example, hog slaughter alone will provide only a rough indication of hog prices. A more satisfactory model would also incorporate other independent variables, such as broiler slaughter. The *multiple regression equation* is a straightforward extension of simple regression and describes the linear relationship between the dependent variable and two or more independent variables.

The meaning of *linear,* which might not be intuitively obvious beyond the two-dimensional case, is that all the variables are of the first degree and are combined only by addition or subtraction. For example, in terms of Z as a function of X and Y, $Z = 2X + Y + 3$ is a linear equation, while $Z = X^2 + 2y^2 + 4$, $Z = XY$, and $Z = \log X + \log Y$ are nonlinear equations. A basic characteristic of a linear equation is that a one-unit change in an independent variable will result in a *constant* magnitude change in the dependent variable, regardless of the independent variable value. In other words, in a linear equation, the slope in each dimension is constant. When there are only two variables, as is the case in simple regression, the linear equation can be depicted by a straight line. When there are three variables, the linear equation can be represented by a plane in three-dimensional space. Linear equations involving more than three variables can no longer be simply represented in three-dimensional Euclidean space.

As in the simple regression case, regression analysis is only appropriate if the relationship between the variables is approximately linear. This is not as strict a limitation as it may sound, since many nonlinear equations can be transformed into linear equations.

The general form of the multiple regression equation is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_k X_k + e$$

The preceding equation represents the unknown population or true regression. The general form of the fitted regression is

$$Y = a + b_1 X_1 + b_2 X_2 \ldots b_k X_k$$

where a, b_1, b_2, . . . , b_k are chosen so as to minimize the sum of the squared residuals. A regression coefficient b_i can be interpreted as follows: If all other independent variables are held constant, a one-unit change in X_i will cause the dependent variable Y to change by b_i. A residual can still be interpreted as the difference between an observed value of the dependent variable (Y_i) and the fitted value (Y_i). Regardless of how many variables there are in the equation, the residuals still represent a single-dimensional difference.

These concepts can perhaps be clarified by considering a practical example. Assume that we have derived a regression equation relating hog prices to hog slaughter and broiler slaughter:

$$Y = a_1 + b_1 X_1 + b_2 X_2$$

where Y = deflated hog prices
X_1 = hog slaughter
X_2 = broiler slaughter

This relationship is depicted in Figure D.1. Each combination of values for X_1 and X_2 will fix a location in the (X_1, X_2) plane. The regression equation will indicate the value of Y (the height of the y axis) at that point. In other words, the regression equation defines the price level that corresponds to any combination of hog and broiler slaughter.

For any given value of X_2 (broiler slaughter), increases in hog slaughter will reduce Y (hog prices) at a constant rate. Similarly, for any given value of X_1 (hog slaughter), increases in broiler slaughter will reduce Y. Thus, as can be seen, the highest prices will occur when hog slaughter and broiler slaughter are low, and the lowest prices when these explanatory variables are high.

Each observation will represent a combination of values for X_1 and X_2 and the corresponding value of Y during the given period. Of course, the actual observed Y_i values, which are indicated by solid dots in Figure D.1, will rarely fall exactly on the plane. The vertical distance between any point and the plane (i.e., the difference between the actual hog price and the price predicted by the regression plane) is the residual and is indicated by an arrow. As in the simple regression case, the residual is measured along a single axis (y). The regression procedure will specify the values for a, b_1, and b_2, which will minimize the sum of the squares of these residuals.

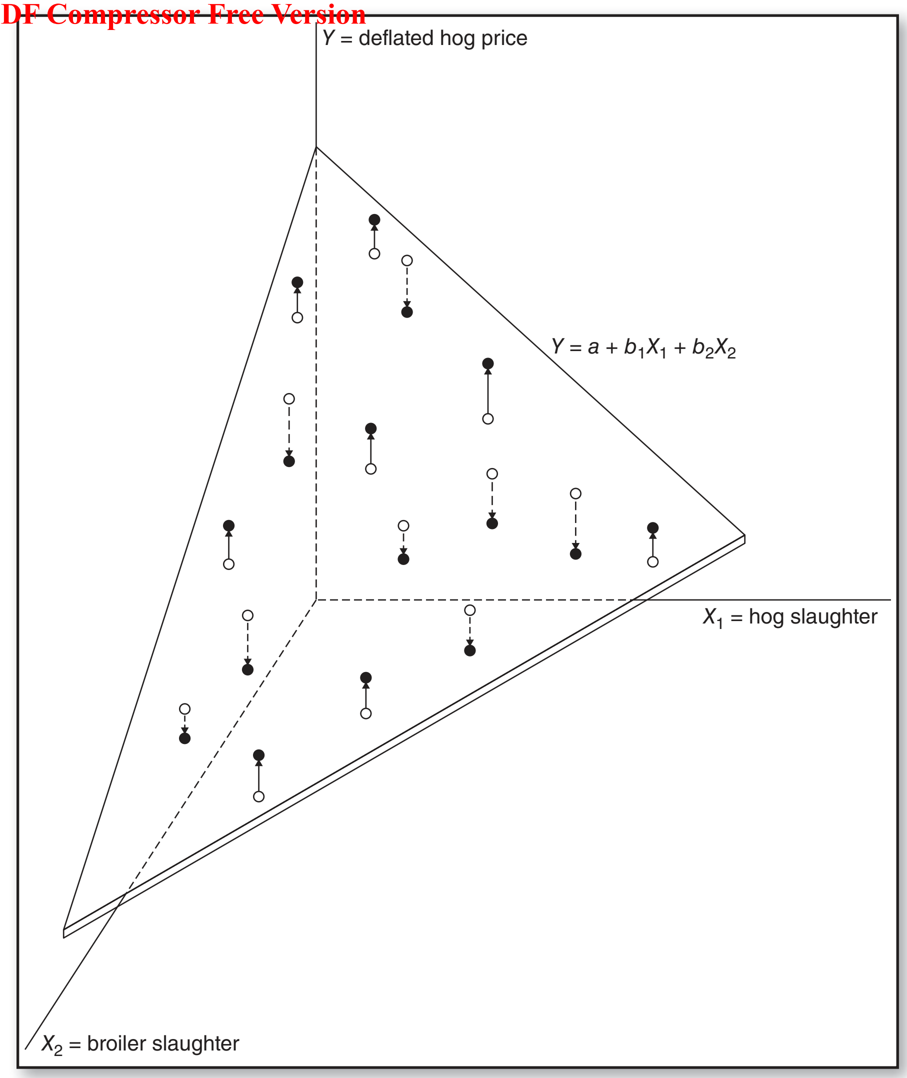


FIGURE D.1 Scatter of Observed Points About the Regression Plane

We deliberately avoid indicating the computational formulas for deriving the regression coefficients, or for that matter, any of the statistics for the multiple regression case. The reason for this is that multiple regression computations are far too cumbersome to be reasonably considered without the aid of a computer. The assumptions underlying multiple regression are completely analogous to those detailed for the simple regression model. In the multiple regression case, there is one additional assumption: that there is no linear relationship among any two or more independent variables. If this assumption is not met, it will lead to a problem called *multicollinearity*.

Applying the *t*-Test in the Multiple Regression Model

The t values in the multiple regression output can be used to evaluate the significance of the regression coefficients. The t values from any standard computer regression analysis assume the hypothesis being tested is that the regression coefficient equals zero. In this case, the t value for b_i is provided by

$$t = \frac{b_i}{\text{s.e.}(b_i)}$$

This relationship can be expressed as follows:

$$\text{T-STAT} = \frac{\text{COEFF}}{\text{ST ER}}$$

where T-STAT = t value for given regression coefficient
COEFF[1] = value of given regression coefficient
ST ER = standard error of given regression coefficient (not to be confused with the standard error of the regression, SER, which is described in the next section)

Frequently, these statistics will be listed in three adjacent columns with each row providing statistics for the constant term or specified regression coefficient. The interpretation of the t statistic is identical to the simple regression case. The t value indicates the number of standard deviations between the indicated coefficient value and the true population coefficient if the latter were actually equal to zero. The higher the t value, the more significant the regression coefficient. To use the t table (Table B.4), one would check the row corresponding to $n-k$ degrees of freedom (*df*) where n equals the total number of observations and k equals the number of variables in the equation. Roughly speaking, t values above 2.0 are clearly significant and indicate that the given independent variable should be retained in the model. It should be noted that the t value does not definitively prove significance; rather it establishes significance at a specific probability level. The higher the t value, the more unlikely a regression coefficient would be assumed to be significant when it is not.

What if the t statistic is less than the t value at the 0.05 level of significance (e.g., less than 1.812 for a one-tailed test with 10 *df*)?[2] Here there is no clear-cut answer. The choice depends on the priorities of the analyst. If she is more concerned about retaining an insignificant variable in the model, she might lean to dropping any variable whose regression coefficient is not significant at the 0.05 level. On the other hand, if she is more concerned about deleting a meaningful variable, she would retain the variable unless the t value was very low.

A reasonable criterion is that any theoretically meaningful variable with a t value greater than 1.0 should usually be retained,[3] although it should be noted that many analysts prefer to use a cutoff

[1] Frequently also called VALUE.

[2] It is assumed that the regression coefficient has the anticipated sign (e.g., negative for hog slaughter in an equation in which hog prices are the dependent variable). Situations in which the sign is opposite to expectations are discussed later in this appendix.

[3] There is a special meaning to t values > 1.0. It has been demonstrated that if explanatory variables are retained if their t-value > 1.0 and deleted otherwise, the "Corrected R^2" (which is discussed later) will be maximized.

level of 2.0. The key words are *theoretically meaningful*. A low t value does not contradict the assumed relationship between the dependent and explanatory variable. Remember, a t value below the level of statistical significance does not indicate that the independent variable is not meaningful in explaining the dependent variable. It only means that its significance has not been demonstrated at the desired probability level. As long as the variable has the anticipated sign, the results are still consistent with theoretical expectations, albeit the relationship is not as strong as would be desired. Furthermore, even a t value of 1.0 would still be significant at the 0.20 level (i.e., 80 percent probability) for any regression equation in which $df > 2$. Variables with t values below 1.0 should usually be dropped.

There is one exception to the decision process just detailed. Occasionally, the analyst might try including all the independent variables she believes should significantly affect the dependent variable, only to find that the resulting regression equation is disappointing. At this point, in desperation she might try a variety of independent variables in the hopes that perhaps one or more of these are significantly related to the dependent variable. Such a method could be termed a "shotgun" or "kitchen sink" approach and is not recommended unless all theoretically plausible variables have been exhausted. In any event, in this case one should apply stricter requirements for retaining a variable. First, a two-tailed rather than one-tailed test should be used (see section "Testing the Significance of the Regression Coefficients" in Appendix C). Second, variables with t values below the 0.05 level of significance should be rejected. In fact, one can argue that a more restrictive significance level should be adopted (e.g., 0.01), since the probability of accepting a meaningless variable increases with the number of variables tested.

Thus far, we have assumed that a theoretically chosen variable has the correct sign. However, in equations with many variables, a coefficient with the wrong sign is not uncommon. Such an occurrence usually indicates the presence of multicollinearity—a linear dependence between two or more explanatory variables. (A discussion of how to handle such variables is presented in the section on multicollinearity in Appendix E.) At this point, suffice it to say that the t values of such variables are usually irrelevant.

Standard Error of the Regression

The standard error of the regression (SER) is a measure of the unexplained variation. The definition of the SER is almost totally analogous to the simple regression case. The only difference is that the sum of the squared residuals is divided by the appropriate degrees of freedom, instead of $n-2$. Thus, for the more general multiple regression case, the SER could be expressed as

$$\text{SER} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n - k}$$

where k = number of parameters in equation (which is equal to the number of independent variables plus 1, assuming there is a constant term in the equation). Note in the simple regression case that $k = 2$.

As in the simple regression case, the % SER is equal to the SER divided by $\overline{Y}$. Where appropriate (see Appendix C), the % SER may be more convenient to use, because it is stated in a form that is intuitively meaningful.

Confidence Intervals for an Individual Forecast

In the multiple regression case, the calculation of a confidence interval for an individual forecast is somewhat complicated. As a simplification, the confidence interval can be calculated for the situation in which all of the independent variables are equal to their means. In this specialized case, the formula for the confidence interval would be analogous to the simple regression case in which $X = \overline{X}$:

$$\hat{Y}_f - t \cdot s\sqrt{1 + \frac{1}{n}} < Y_f < \hat{Y}_f + t \cdot \sqrt{1 + \frac{1}{n}}$$

where s = SER

t = t value at specified level of significance for the given degrees of freedom

This represents a minimum confidence interval, and the further removed the independent variables are from their respective means, the wider the actual confidence interval.

R^2 and Corrected R^2

The term R^2 is the multiple regression counterpart of r^2 and is defined in exactly the same way. Thus, the entire discussion related to r^2 in Appendix C applies here as well and need not be duplicated.

In the multiple regression case, it is important to realize that the addition of another independent variable can only increase R^2. Remember that R^2 is the ratio of explained variation to total variation. The introduction of a new variable will not affect total variation, and it can only increase explained variation. Even the introduction of a totally irrelevant variable will probably result in a small increase in explained variation. For example, it is a safe bet that adding a variable for the number of ducks in Belgium would increase the R^2 of a regression equation for forecasting U.S. interest rates.

The point that the addition of a meaningless explanatory variable will raise R^2 is more than an esthetic consideration. Recall that each additional variable will decrease the degrees of freedom by 1, thereby reducing the significance of the equation on the basis of other measures such as the t-test and SER, all else being equal. For this reason, it is desirable to modify the R^2 measure so that it is penalized for the addition of irrelevant variables. This alternative measure is called the *Corrected* R^2 (CR^2), or sometimes the *Adjusted* R^2. The problem with R^2 is that it is based on *variation*, which does not account for the number of degrees of freedom. The CR^2 avoids this defect, because it is based on variance. The variance is simply the variation divided by the number of degrees of freedom. It will be recalled that R^2 can be defined as[4]

$$R^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

[4] The formulas for R^2 and r^2 are identical.

We now define CR^2 as

$$CR^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}}$$

where

$$\text{Variance} = \frac{\text{variation}}{df}$$

Thus,

$$CR^2 = 1 - \frac{\dfrac{\sum_{i=1}^{n}(Y_1 - \hat{Y}_i)^2}{n-k}}{\dfrac{\sum_{i=1}^{n}(Y_1 - \bar{Y})^2}{n-1}}$$

The numerator of the ratio term is based on n observations, but there are k constraints in finding the regression line used to calculate Y_i. Thus $df = n - k$. The denominator is also based on n observations, but there is only one constraint, $\bar{Y}$; thus $df = n - 1$. The preceding equation can be rewritten as

$$CR^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k}$$

As is readily apparent in this form of the equation, when n is large relative to k, the CR^2 will almost equal R^2.

Typical regression runs will provide the CR^2 (corrected R square or adjusted R square) along with R^2 (R square). As a general rule, the CR^2 is a more useful measure for comparing different regression equations for the same dependent variable.

F-Test

Whereas the t distribution is used to test the significance of the individual regression coefficients, the F distribution is used to test the significance of the regression equation as a whole. In other words, the F statistic tests the hypothesis that none of the regression coefficients is significant. The F statistic can be expressed as

$$F = \frac{\text{explained variance}}{\text{unexplained variance}}$$

Note that the *F* value is based on variance, not variation. Once again, variance = variation + *df*.

$$F = \frac{\dfrac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{k-1}}{\dfrac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2} \cdot \frac{n-k}{k-1}$$

The degrees of freedom for the explained variance = $k - 1$, since *k* values are employed in defining the regression line used to calculate $\hat{Y}_i$, but one *df* is lost because of the constraint imposed by $\overline{Y}$. As for the unexplained variance, there are *n* observations, but *k* constraints are imposed in finding the regression line upon which Y_i is based. Recalling the alternative definitions for R^2, we can re-express *F* as[5]

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n-k}{k-1}$$

The appropriate degrees of freedom will be specified in the notation for the *F* statistic. For example, *F* (2/8) = 23.5 indicates an *F* value for a regression equation in which $k - 1 = 2$ and $\text{n} - k = 8$. To check for significance, the *F* statistic is compared to the listed values in the *F* table for the corresponding number of degrees of freedom. For example, checking Table D.1, it can be determined that at the 0.01 level of significance, *F*(2/8) = 8.65; thus, a value of 23.5 would be significant.

In practice, the *F*-test is not particularly critical, since it will almost invariably prove significant. This should not be surprising, because the *F*-test checks whether all the regression coefficients combined have any predictive value—a very weak criterion. In any case, for comparisons of regression equations with the same dependent variable, higher *F* values would indicate a better model (assuming none of the regression assumptions are violated). However, similar information could be gathered by comparing CR^2 values.

Analyzing a Regression Run

Table D.2 presents the results for a sample regression run. At this juncture, most of Table D.2 should be comprehensible. However, it may be helpful to interpret the key statistics of this table.

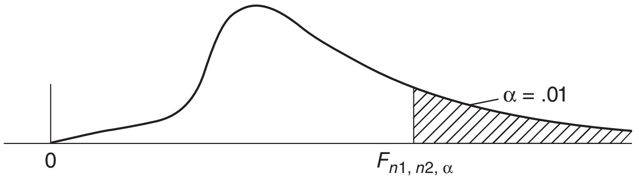
1. The regression equation is $Y = 49.06899 - 1.07049\ (X1) + 0.35775\ (X2)$. To get a point forecast for *Y*, one would merely plug in the estimated values of *X*1 and *X*2. For example, if $X1 = 20$

[5] $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = R^2 \cdot \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$ and $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = (1 - R^2)\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

TABLE D.1 *F* Distribution

Values of $F_{n1,n2,\alpha}$ on the $F(_{n1,n2,\alpha})$-distribution

$\Pr\{F(_{n1,n2})\text{-variable} \geq F_{n1,n2,\alpha}\} = \alpha = 0.01$



α = .01

***F*(n_1, n_2)-distribution**

	n_1 (numerator *df*)								
n^2 (denominator *df*)	**1**	**2**	**4**	**6**	**8**	**10**	**12**	**24**	**∞**
	$[t_{n2,.005}]^2$	**Values of $F_{n1,n2,\alpha}$**							
1	4,052	5,000	5,625	5,859	5,982	6,056	6,106	6,235	6,366
2	98.50	99.00	99.25	99.33	99.37	99.40	99.42	99.46	99.50
3	34.12	30.82	28.71	27.91	27.49	27.23	27.05	26.60	26.13
4	21.20	18.00	15.98	15.21	14.80	14.55	14.37	13.93	13.46
5	16.26	13.27	11.39	10.67	10.29	10.05	9.89	9.47	9.02
6	13.75	10.92	9.15	8.47	8.10	7.87	7.72	7.31	6.88
7	12.25	9.55	7.85	7.19	6.84	6.62	6.47	6.07	5.65
8	11.26	8.65	7.01	6.37	6.03	5.81	5.67	5.28	4.86
9	10.56	8.02	6.42	5.80	5.47	5.26	5.11	4.73	4.31
10	10.04	7.56	5.99	5.39	5.06	4.85	4.71	4.33	3.91
11	9.65	7.21	5.67	5.07	4.74	4.54	4.40	4.02	3.60
12	9.33	6.93	5.41	4.82	4.50	4.30	4.16	3.78	3.36
13	9.07	6.70	5.21	4.62	4.30	4.10	3.96	3.59	3.17
14	8.86	6.51	5.04	4.46	4.14	3.94	3.80	3.43	3.00
15	8.68	6.36	4.89	4.32	4.00	3.80	3.67	3.29	2.87
20	8.10	5.85	4.43	3.87	3.56	3.37	3.23	2.86	2.42
25	7.77	5.57	4.18	3.63	3.32	3.13	2.99	2.62	2.17
30	7.56	5.39	4.02	3.47	3.17	2.98	2.84	2.47	2.01
40	7.31	5.18	3.83	3.29	2.99	2.80	2.66	2.29	1.80
60	7.08	4.98	3.65	3.12	2.82	2.63	2.50	2.12	1.60
120	6.85	4.79	3.48	2.96	2.66	2.47	2.34	1.95	1.38
∞	6.63	4.61	3.32	2.80	2.51	2.32	2.18	1.79	1.00

Source: Abridged from Table 18 of Pearson and Hartley, *Biometrika Tables for Statisticians,* Third Edition, Volume 1, 1976, with kind permission of the *Biometrika* Trustees (http://biomet.oxfordjournals.org).

The diagram, and this presentation of the table, are taken from Table 4b of S. R. Searle, *Linear Models* (New York, NY: John Wiley & Sons, 1997). Copyright © 1997 by John Wiley & Sons; reprinted by permission.

TABLE D.2 Sample Regression Run Results

Variable	Coefficient	Standard Error	*T*-Stat	Mean
CONSTANT	49.06899	9.67267	5.07	41.16071 (dependent variable)
X1	−1.07049	0.23464	−4.56	21.00714
X2	0.35775	0.13400	2.67	40.75357
Observation	**Actual**	**Fitted**	**Residual**	**% Deviation**
1	35.90000	35.25925	0.64075	1.82
2	52.70000	54.54910	−1.84910	−3.39
3	46.30000	50.74680	−4.44680	−8.76
4	34.20000	36.98609	−2.78609	−7.53
5	51.30000	46.15574	5.14426	11.15
6	44.20000	44.02220	0.17780	0.40
7	33.90000	29.70675	4.19325	14.12
8	31.30000	30.54304	0.75696	2.48
9	31.70000	32.74207	−1.04207	−3.18
10	29.90000	31.58795	−1.68795	−5.34
11	51.10000	49.76981	1.33019	2.67
12	56.10000	51.62468	4.47532	8.67
13	43.90000	45.56465	−1.66465	−3.65
14	33.7500	36.99187	−3.24187	−8.76

Y = CONSTANT + C1 · X1 + C2 · X2

RSQ = 0.8953   SER = 3.2338   F(2,11) = 47.0

RSQC = 0.8762   % SER = 7.86   DW = 1.69

and $X2 = 40$, the predicted Y value would be 41.969. In practice, it will be more convenient to use mnemonic symbols for the variables instead of Y, $X1$, and $X2$.

2. $R^2 = 0.8953$, which means that $X1$ and $X2$ explain 89.53 percent of the total variation in Y. The CR^2, which is adjusted downward for lost degrees of freedom, is 0.8762.
3. SER = 3.2338. This would be a key figure of merit in comparisons with alternative models. The SER could also be used to construct a crude confidence interval for an individual forecast based on the assumption that all the independent variable values are equal to their respective means. This confidence interval would be[6]

$$\hat{Y}_f - t \cdot s\sqrt{1 + \frac{1}{n}} < Y_f < \hat{Y}_f + t \cdot s\sqrt{1 + \frac{1}{n}}$$

[6] See the section, "Confidence Intervals for an Individual Forecast."

where s = SER = 3.23

$n = 14$

$t = 2.201$ (t value for two-sided test at 0.05 level of significance for 11 *df*)

$$\hat{Y}_f - 2.201\,(3.23)\,(1.0351) < Y_f < \hat{Y}_f + 2.201\,(3.23)\,(1.0351)$$
$$\hat{Y}_f - 7.3588 < Y_f < \hat{Y}_f + 7.3588$$

Using the point estimate for Y_f of 41.969 derived in 1, the 95 percent confidence interval would be

$$34.610 < Y_f < 49.328$$

This would mean a 95 percent probability the actual value will fall in the stated range if the forecast is based on independent variables equal to their respective means. Of course, this will never be the case. Consequently, the actual confidence interval will always be wider. Nevertheless, given this understanding, the simplified confidence interval provides at least a rough sense of the potential variability of the forecast.

4. The %SER $=$ SER $\div \overline{Y}$. The %SER provides a figure that is intuitively meaningful and can be used instead of the SER if all model comparisons involve the same dependent variable.
5. The F value for 2 *df* in the numerator and 11 *df* in the denominator = 47.0, which is well above the listed F value of 7.21 at the 0.01 significance level. Again, the F test will almost invariably verify that the equation is significant.
6. DW stands for Durbin-Watson, a measure discussed in Appendix E..
7. The t statistic is equal to the coefficient values divided by the respective standard error. In this example, all the coefficients are significant (t value at the 0.05 level of significance for a one-tailed test with 11 *df* is 1.796).
8. The Actual column in Table D.2 lists the actual observations of Y, and the Fitted (also called Predicted) column lists the corresponding values indicated by the regression equation. The difference between these two figures for each observation is listed in the Residual column. The Percent Deviation column is equal to the residual divided by the fitted value. (In some cases, the residuals are normalized by dividing the residuals by the SER—that is, expressing residuals in standard deviation units. Residuals normalized in this manner are termed *standardized residuals* or *standard residuals* and are discussed in Appendix E.)

Thus far, we have only discussed the meaning of the overall summary statistics for the regression equation. As will be detailed in Appendix E, the individual residual values also contain extremely important information and should be carefully analyzed.

Analyzing the Regression Equation

It is a test of true theories not only to account for but to predict phenomena.

—William Whewell

Outliers

An outlier is an observation with a large residual—that is, a large deviation between the observed value and fitted value. Outliers reflect one of the following conditions:

1. An error in collecting or manipulating the data for the given point.
2. The existence of a significant extraneous causal factor that only affected the outlier(s).
3. The omission of an important explanatory variable from the equation.
4. A structural flaw in the model.

The presence of outliers indicates a deficiency in the model. After verifying that an outlier is not the result of error, one should try to identify possible factors responsible for the aberrant behavior. If the outlier can be explained by a missing variable that affected all observations, then this variable should be included in the equation. If, however, the outlier was a consequence of an isolated event that is not expected to reoccur, then it should be viewed as an unrepresentative point, and the regression should be rerun with the outlier deleted. This recalculation is important, since the method of least squares used to derive the regression coefficients will give greater weight to outliers. Thus, one or two such points could seriously distort the regression equation fit. However, unless the isolated causes of the outlier have been identified, one should avoid the temptation of deleting such points simply because it will improve the regression fit.

The Residual Plot

The scatter diagram, such as the one depicted in Figure A.4, is of limited use in detecting outliers, since it can only be applied in the simple regression case. The residual plot provides a graphic technique for detecting outliers that is as easy to apply in multiple regression as it is in simple regression. In constructing residual plots, it is more convenient to use standardized residuals rather than the actual residuals, which vary widely from case to case. The *standardized residual* for the *i*th observation is defined as

$$sr_i = \frac{Y_i - \hat{Y}_i}{s}$$

where

$$s = \text{SER} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n - k}}$$

where n = number of observations
k = number of parameters (which is equal to the number of independent variables plus 1, assuming there is a constant term in the equation)

In effect, the standardized residual can be thought of as indicating how many standard deviations a residual is from the assumed residual mean of zero. If the regression assumptions are valid, the standardized residuals should be randomly distributed and fall primarily in a range between +2 and −2. One great advantage of using standardized residuals is that the interpretation of a residual plot is the same for all types of data. There are three basic types of residual plots:

1. sr_i versus the fitted Y values ($\hat{Y}_i$).
2. sr_i versus the independent variables (in multiple regression, there can be one such plot for each independent variable).
3. sr_i versus time (i.e., the sr_i values are plotted in time order).

The preceding plots are available on some computer packages. However, even when they are not, they can easily be constructed from the printout of residual values. Since virtually all applications of regression analysis in forecasting futures involve time series data, the third type of residual plot will normally be the most useful.

Figure E.1 provides an example of a residual plot plotted against time. As is readily visible, 2011 and 2014 are outliers. Also note the predominance of negative residuals in the remaining years—a consequence of the two positive outliers pulling up the regression line. (Remember, the least-squares procedure will tend to place greater weight on outliers.)

One essential attribute of the residual plot is that it can be applied just as easily when a model contains three or more variables, a situation in which a scatter diagram cannot be constructed. Perhaps even more importantly, the residual plot can be used to check for autocorrelation. Before turning to this critical application of the residual plot, it is first necessary to discuss autocorrelation and the most commonly used method for its detection, the Durbin-Watson (DW) statistic.

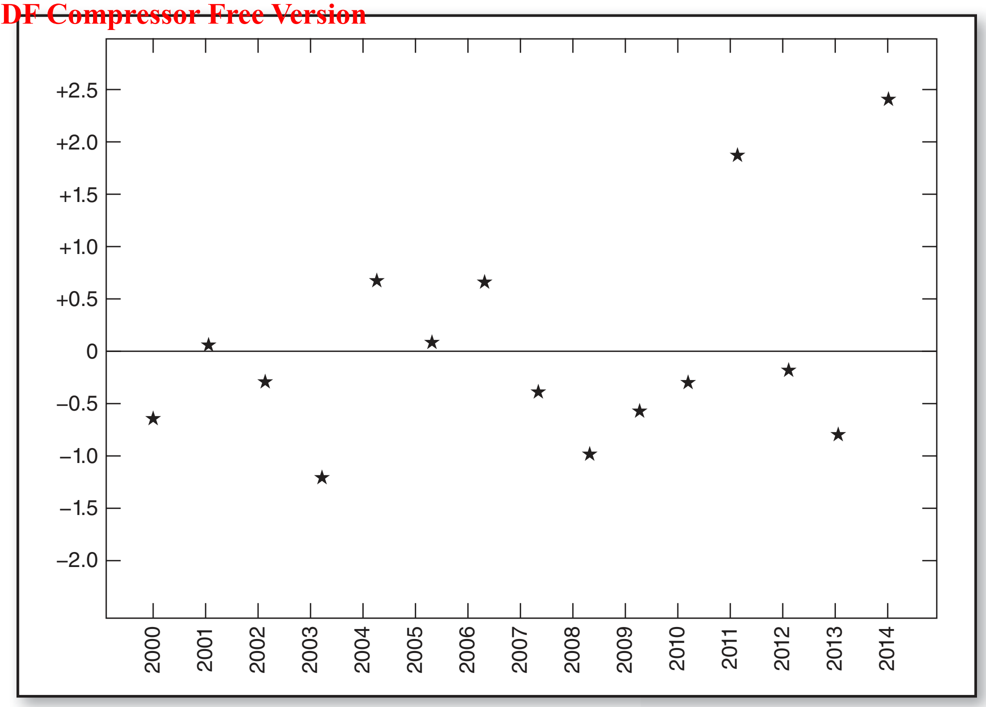


FIGURE E.1 Standardized Residuals Plotted Against Time

Autocorrelation Defined

Autocorrelation refers to the situation in which the error terms are correlated. The existence of autocorrelation indicates that there is a pattern in the data that has not yet been captured by the explanatory variables. For this reason, the presence of autocorrelation suggests that the model is still inadequate. Furthermore, it should be recalled that one of the basic assumptions underlying the statistics of regression analysis is that the error terms are randomly distributed. If autocorrelation exists, then the standard error of the coefficients and the SER may all be severely understated. Thus, the normal tests of significance may yield a very distorted picture of the equation's precision.

The Durbin-Watson Statistic as a Measure of Autocorrelation

Autocorrelation refers to a linear dependency in the error terms. In the simplest case, the error terms are correlated to the error terms of the preceding period. This type of condition is called *first-order autocorrelation* and is reflected by the DW statistic, which is a standard measure included in the regression summary printout.

Although ideally a test for autocorrelation would be based on the population error terms, these are unavailable. The DW test is based on the residuals (denoted here by $\hat{e}_t$) and defined as:

$$DW = \frac{\sum_{t=2}^{n} \left(\hat{e}_t - \hat{e}_{t-1} \right)^2}{\sum_{t=2}^{n} \hat{e}_t^2}$$

where $\hat{e}_t$ = residual value in time period t

$\hat{e}_t - 1$ = residual value in time period immediately preceding period t

The following approximate relationship, however, is more useful for an intuitive explanation of the Durbin-Watson statistic:

$$DW \approx 2(1 - r)$$

where

$$r = \frac{\sum_{t=2}^{n} \hat{e}_t \cdot \hat{e}_{t-1}}{\sum_{t=2}^{n} \hat{e}_t^2}$$

If there is no first-order autocorrelation, the positive values for the product $\hat{e}_t \cdot \hat{e}_{t-1}$ will tend to offset the negative values, and $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ should approximate zero. In this case, r will approach zero and DW will approach 2. If adjacent residuals are positively correlated, then $\hat{e}_t \cdot \hat{e}_{t-1}$ and $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ will tend to be positive. The greater the correlation between adjacent residuals, the more positive $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ will be. In the extreme, $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ will approach $\sum \hat{e}_t^2$, causing r to approach 1 and DW to approach 0. Similarly, if the adjacent residuals are negatively correlated (i.e., positive residuals tending to follow negative residuals), $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ will be negative. If the negative correlation is extreme, $\sum \hat{e}_t \cdot \hat{e}_{t-1}$ will approach $-\sum \hat{e}_t^2$, causing r to approach -1 and DW to approach 4. In summary, the *DW* has a possible range between 0 and 4. Values near 2 indicate no first-order autocorrelation, values significantly below 2 indicate positive autocorrelation, and values significantly above 2 indicate negative autocorrelation.

Table E.1 contains a list of DW values that can be used to test for autocorrelation at the 0.05 level of significance. The appropriate values will depend on the number of observations, n, and the number of independent variables in the regression equation, k. Note that unlike the other tests discussed so far, there are two values listed for each category. In a test for positive autocorrelation, the interpretation would be as follows (for negative autocorrelation use 4 − DW instead of DW):

1. $DW < d_L$ Positive autocorrelation exists
2. $DW > d_U$ No positive autocorrelation exists
3. $d_L < DW < d_U$ Test is inconclusive

TABLE E.1 The Distribution of Durbin-Watson *d* 5 Percent Significance Points of d_l and d_u

	$k = 1$		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Note: *DW* values below d_L indicate that positive autocorrelation exists; values above d_U indicate that no positive autocorrelation exists. *DW* values between d_L and d_U are inconclusive. To test for negative correlation use 4 − *DW* instead of *DW.*

Source: S. Chatterjee and B. Price, *Regression Analysis by Example*, 3rd ed. (New York, NY: John Wiley & Sons, 1999). Copyright © 1999 by John Wiley & Sons; reprinted with permission.

For example, assume we are testing a regression equation with 18 observations and three variables. Positive autocorrelation would be indicated if $DW < 0.93$, no autocorrelation if $DW > 1.69$, and the test would be inconclusive if $0.93 < DW < 1.69$. The test for negative autocorrelation would be analogous, using $4 - DW$ instead of *DW.*

A routine check of summary statistics for a regression equation should include the DW. A particularly low or high DW would indicate a definite need for further analysis and model modification. However, it should be emphasized that even a perfect DW value (2.0) does not guarantee that autocorrelation does not exist. The DW only tests for first-order autocorrelation. If the interrelationship between the error terms is more complex, the DW might not pick it up. For this reason, it is also advisable to check a residual plot for autocorrelation. Furthermore, as will be illustrated in subsequent sections, the residual plot can also be used to provide important clues for improving the regression model.

The Implications of Autocorrelation

The presence of a pattern in the residuals suggests a potential inadequacy in the regression equation. Specifically, autocorrelation may reflect one of the two following flaws:

1. The omission of significant explanatory variables in the regression equation.
2. The use of the linear regression method to describe a nonlinear relationship between the dependent and independent variables.

If the autocorrelation is due to one of these factors, it is clear why autocorrelation is undesirable. These conditions indicate that a better model can be constructed, either by adding variables to the equation or by trying different functional relationships. However, even when this is not the case, an equation that exhibits autocorrelation is still undesirable, because the violation of the assumption that the error terms are randomly distributed will lead to distortions.[1] For this reason, as a last resort, transformations designed to remove the autocorrelation should be considered.

To summarize, the *DW* and residual plot should be checked for autocorrelation. If residuals are found to be correlated, the following steps should be taken:

1. Try to find any significant variables that may have been omitted from the equation.
2. If all feasible variables have been tried and autocorrelation still exists, experiment to see whether alternative functional forms (other than the linear form assumed in the regression procedure) are more appropriate.
3. If both of the preceding steps are unsuccessful, transformations to remove autocorrelation might be tried.

[1] If autocorrelation exists, the standard regression approach, which is called *ordinary least squares* (*OLS*), will still yield unbiased estimates (i.e., estimates that on average will equal the population parameters). However, the estimates will no longer be *efficient* (i.e., they will not be the minimum variance estimates). Even worse, the standard error estimates of the regression coefficients and the equation as a whole may be severely understated. Consequently, the true confidence interval may be much wider than suggested, and the regression equation may be too imprecise to be used for forecasting.

The first of these steps will be illustrated by an example in the next section. Methods to address the second two steps are discussed in the addendum to this appendix.

Missing Variables and Time Trend

A pattern in the residual plot (or the presence of autocorrelation) can be viewed as an indication that significant explanatory variables are missing from the equation. For example, Figure E.2 shows the residual plot for the simple regression model of the average December hog futures price during July–November as a function of per capita June–November hog slaughter. Note the obvious nonrandom distribution of the residuals: There seems to be a definite trending pattern in the residuals with large negative values predominating in the earlier years and large positive values predominating in the later years.

Given this strong trending pattern in the residuals, we add a time trend as one of the explanatory variables. A time trend is simply a set of successive integers. Normally, the first observation would be assigned a value of 1, the second a value of 2, and so on. However, since the regression model is linear, any set of consecutive integers would serve equally well.

It is not surprising that the fitted values of the original equation tend to be too high in the earlier years and too low in the later years since our model used nominal rather than deflated prices. The reader may well wonder why we didn't first change the model by using deflated prices instead of adding a time trend. In fact, this alternative approach is entirley reasonable as the first change to try,

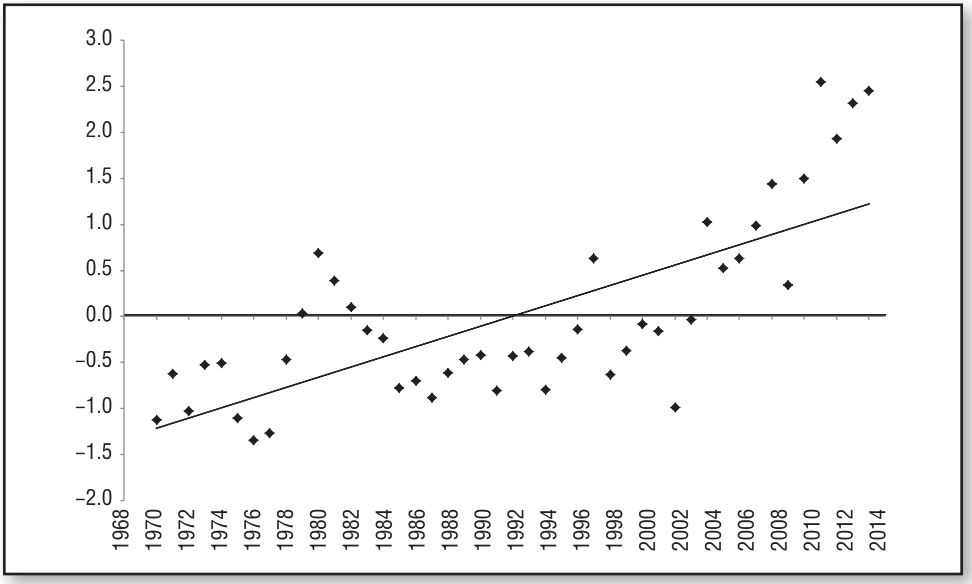


FIGURE E.2 Standardized Residuals for Average Price of December Hog Futures (July–November) vs. June–November Hog Slaughter

and we did run this model (not shown), but the results were substantially inferior to the model that added a time trend.

Table E.2 compares the summary statistics of this two-independent-variable regression equation with those of the original one-independent-variable model. Note the dramatic improvement in all the summary statistics and the significance of the time trend as reflected by its *t* statistic. In fact, the time trend is statistically even more significant in explaining hog prices than hog slaughter! In addition, the strong trend evident in the residual plot for the original simple regression (Figure E.2) has been eliminated in the residual plot for the new equation (Figure E.3). Although the trend in the residuals has been eliminated, Figure E.3 still exhibits a non-random pattern. Specifically, the residuals now conform to a broad "U" pattern, with positive residuals predominating in the early and late years and negative residuals predominating in the middle years. The existence of this pattern suggests other significant variables are still missing from the equation.

Next we add per capita broiler slaughter to the model, since poultry is an important competitive meat to pork. Table E.3 compares the key statistics for this new equation (Model 3) with the corresponding values for the first two models. As can be seen, the addition of poultry slaughter provides a large improvement in all the key statistics. For example, the corrected R^2 jumps from 0.66 in Model 2 to 0.82 in Model 3. Moreover, not only is the *t* statistic for broiler slaughter highly significant but the addition of this variable also increases the *t* statistics for the other explanantory variables (hog slaughter and trend). The addition of broiler slaughter to the equation also eliminates the pattern in the residuals. As can be seen in Figure E.4, which shows the residual plot corresponding to Model 3, the scatter of residuals now seems random.

Achieving a random residual plot doesn't necessarily mean the model is complete. It may well be possible to further improve the model by adding other variables. Model 4 in Table E.3 illustrates

TABLE E.2 **Regression Summary Statistics for Hog-Price-Forecasting Models**

Statistic	Model 1: Hog Price vs. Per Capita Hog Slaughter	Model 2: Hog Price vs. Per Capita Hog Slaughter and Trend
R^2	0.21	0.66
CR^2	0.20	0.64
SER	13.95	9.30
%SER	27.2	18.16
F	11.72	40.62
t-stat (constant)	5.19	4.57
t-stat (hog slaughter)	–3.42	–3.12
t-stat (trend)	NA	7.41
t-stat (broiler slaughter)	NA	NA
t-stat (cattle slaughter)	NA	NA

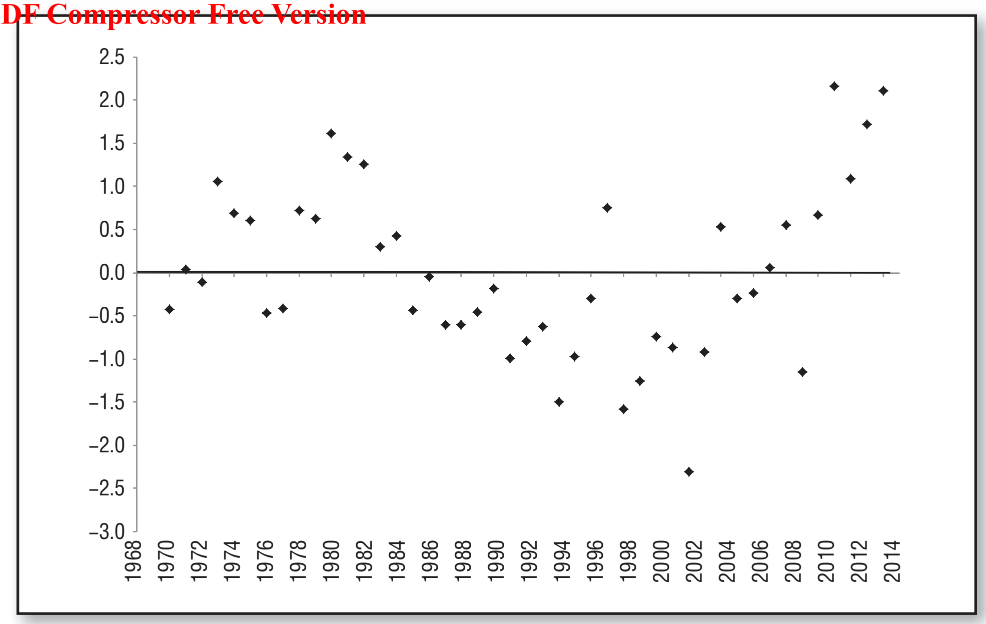


FIGURE E.3 Standardized Residuals for Average Price of December Hog Futures (July–November) vs. June–November Hog Slaughter, and Trend

TABLE E.3 Regression Summary Statistics for Hog-Price-Forecasting Models

Statistic	Model 1: Hog Price vs. Per Capita Hog Slaughter	Model 2: Hog Price vs. Per Capita Hog Slaughter and Trend	Model 3: Hog Price vs. Per Capita Hog Slaughter, Broiler Slaughter, and Trend	Model 4: Hog Price vs. Per Capita Hog Slaughter, Broiler Slaughter, Cattle Slaughter, and Trend
R^2	0.21	0.66	0.84	0.85
CR^2	0.20	0.64	0.82	0.84
SER	13.95	9.30	6.53	6.24
%SER	27.2	18.16	12.76	12.18
F	11.72	40.62	69.51	58.40
t-stat (constant)	5.19	4.57	8.45	6.02
t-stat (hog slaughter)	–3.42	–3.12	–4.39	–5.11
t-stat (trend)	NA	7.41	10.29	5.88
t-stat (broiler slaughter)	NA	NA	–6.64	–7.23
t-stat (cattle slaughter)	NA	NA	NA	–2.22

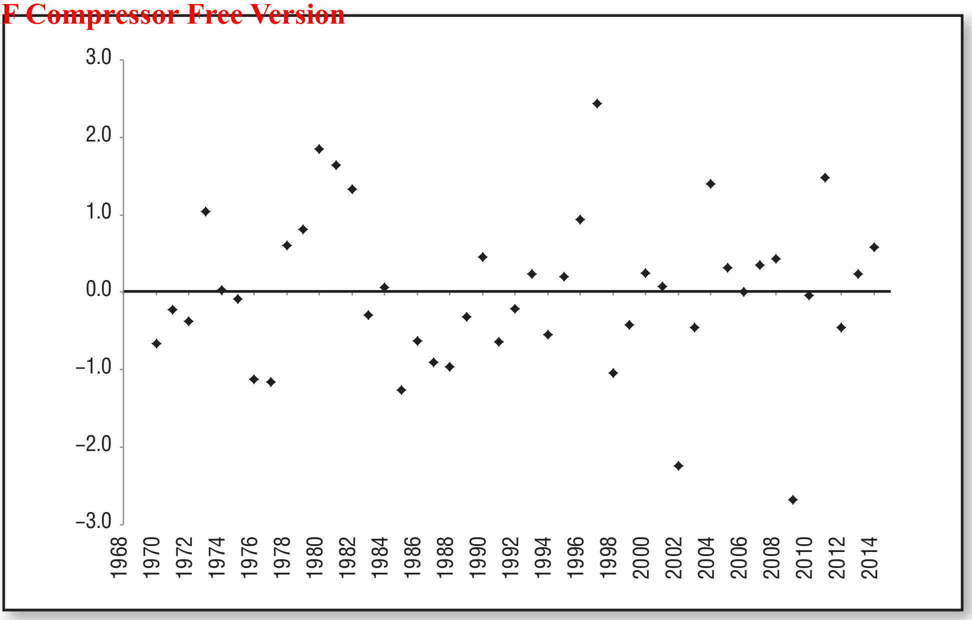


FIGURE E.4 Standardized Residuals for Average Price of December Hog Futures (July–November) vs. June–November Hog Slaughter, Broiler Slaughter, and Trend

one such attempt: adding per capita cattle slaughter to the model on the premise that beef is another competitive meat to pork. The *t* statistic for cattle slaughter is statistically significant and adding this variable modestly increases the corrected R^2 and lowers the SER.

Dummy Variables

In Appendix A we derived a regression equation for forecasting June–November hog slaughter from the prior December–May pig crop. Consider what happens when we attempt to make the equation more general by forecasting hog slaughter during a six-month period from the pig crop of the previous six-month period. In this case, half the observations are those of the original equation, while the other half relate December–May slaughter to the June–November pig crop. Figure E.5 illustrates the residual plot for this equation. We have used two different symbols to distinguish between the residuals for June–November slaughter and the residuals for December–May slaughter. Note the striking pattern of the predominance of positive residuals for June–November slaughter and negative residuals for December–May slaughter. As Figure E.5 dramatically indicates, our equation is missing some important information: the seasonal period of the slaughter forecast. Clearly, we want our equation to distinguish between the two periods. In other words, it is necessary to include a seasonal indicator.

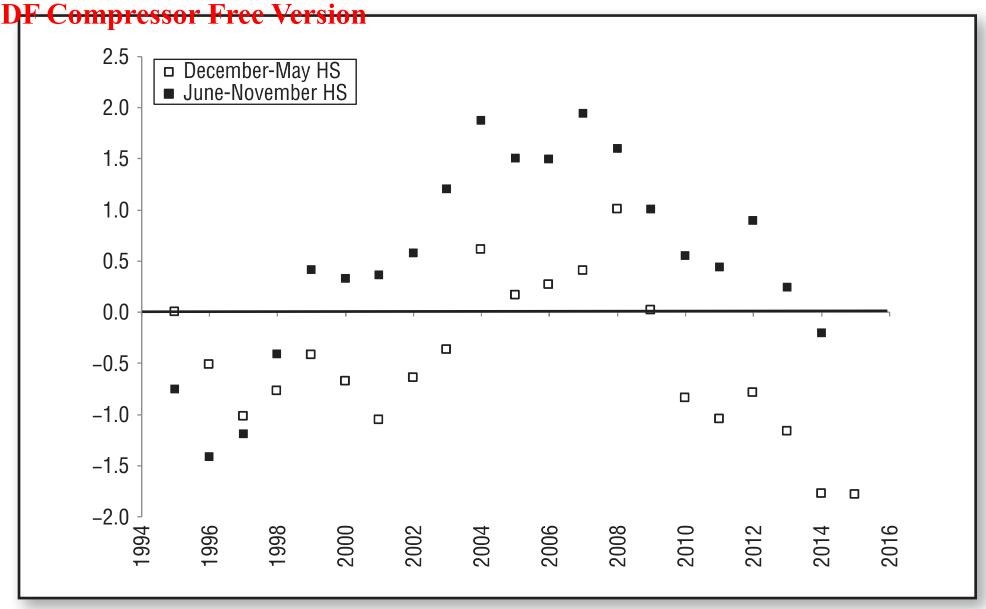


FIGURE E.5 Standardized Residuals for Hog Slaughter vs. Previous Six-Month Pig Crop

A simple method for handling such a situation would be to add a *dummy variable* to the equation, which has a value of 1 for one season and a value of 0 for the other season. The regression equation adding a dummy variable could be written as

$$HS = a + bPC + cS$$

where HS = hog slaughter
PC = pig crop
S = dummy variable, which equals 0 during December–May and 1 during June–November

The dummy variable can be thought of as a switch that is set to off (0) during the base period (December–May) and on (1) during June–November. The dummy variable will have the effect of shifting the intercept by an amount c for the June–November observations. Note that this adjustment will be exactly equivalent to finding two separate equations with the same slope, one for each period. That is, $HS = a + bPC + cS$ for all periods is equivalent to:

$$HS = a_1 + bPC \text{ for December-May slaughter}$$
$$HS = a_2 + bPC \text{ for June-November slaughter}$$

where $a_2 = a_1 + c$

Typically, most users of regression analysis will only employ a dummy variable to shift the intercept, while the slope is assumed to remain constant from period to period. However, in most instances there is no reason to impose an *a priori* restriction that the slopes be equal in different periods. Rather, it seems preferable to begin by using dummy variables for both the intercept and the slopes.[2] Once this full version of dummy variables is run, we can check the t statistics to see which of the dummy variables are significant and then choose the appropriate model accordingly. Thus, in our example, we would begin with:

$$HS = a + bPC + cS + d \cdot S \cdot PC$$

where $S = 0$ during December–May
$S = 1$ during June–November

The form of the equation used will depend on which of the dummy variables proves significant. Some examples:

1. If neither c or d is significant, we would use:

$$HS = a + bPC$$

2. If only c is significant, we would use:

$$HS = a + bPC + cS$$

3. If both c and d are significant, we would use the full-version equation:

$$HS = a + bPC + cS + d \cdot S \cdot PC$$

[2] There are two important exceptions: (1) When one of the periods contains only a few observations, the slope estimate for this period might be unreliable, and it would be better to pool the data in terms of assuming a common slope coefficient for all observations. For example, consider an annual price-forecasting model with 15 observations, three of which coincided with a government program that distorted normal market behavior. In this case, we would definitely only use the dummy variable for the constant term (with the aforementioned three years having a dummy variable value equal to 1), thereby implicitly imposing the restriction of a common slope. The reason for this is that a slope estimate based on only three observations would not be very reliable. This example illustrates one of the advantages of using dummy variables, as opposed to separate equations for each set of observations. (2) When the number of all possible dummy variables is large compared with the number of observations, it may be desirable to conserve degrees of freedom by limiting the number of dummy variables.

Note that in this last case, when both *c* and *d* are significant, the resulting equation is equivalent to the following two separate equations for each period:[3]

$$HS = a + bPC \quad \text{for December–May}$$
$$HS = (a + c) + (b + d)PC \quad \text{for June–November}$$

Why, then, do we not just run separate equations for each period? There are several reasons:

1. By pooling the data, we increase the number of degrees of freedom and add to the statistical reliability of the equation.
2. We do not know beforehand which, if any, of the dummy variables will be significant. The single-equation approach will allow us to eliminate the dummy variables that appear insignificant, thereby providing a better model. In contrast, the two-equation approach is equivalent to automatically assuming that all the dummy variables are significant.
3. In terms of the various tasks of checking alternative models, testing for significance, and forecasting, it is more convenient to have a single equation that is applicable to all periods than a separate equation for each period.
4. As mentioned in footnote 2, there are times when it is definitely preferable to impose slope restrictions—an approach that requires the use of dummy variables.

Since in our example of hog slaughter versus the prior six-month pig crop the dummy variable for the slope is statistically significant, we use the full form of the equation:

$$HS = a + bPC + cS + d \cdot S \cdot PC$$

Figure E.6 shows the residual plot for the regression equation that adds dummy variables for the slope and intercept. Note that the positive bias for June–November residuals and the negative bias for December–May residuals has been eliminated.

The failure to include dummy variables when they are appropriate will bias the regression coefficient estimates. In Figure E.7 we provide a hypothetical example where the points associated with two different periods are best described by best-fit lines with different constant terms. Note how the slope of a single regression equation line without inclusion of dummy variables is biased by the failure to distinguish between the two periods.

Although our example involved only two periods (one period other than the base period), the dummy variable approach can be extended to more period divisions. For example, if we were using a quarterly model, there would be one dummy variable for each quarter other than the base quarter.

$$Y = a + bX + c_1S_1 + c_2S_2 + c_3S_3 + d \cdot S_1 \cdot X + e \cdot S_2 \cdot X + f \cdot S_3 \cdot X$$

[3] Although the intercept and slopes will be identical in the one- and two-equation versions, there is a minor technical difference between the two models. The single equation implicitly assumes a common variance for all periods, while the two-equation version allows for different variances in each period. This difference could theoretically affect the various tests of significance.

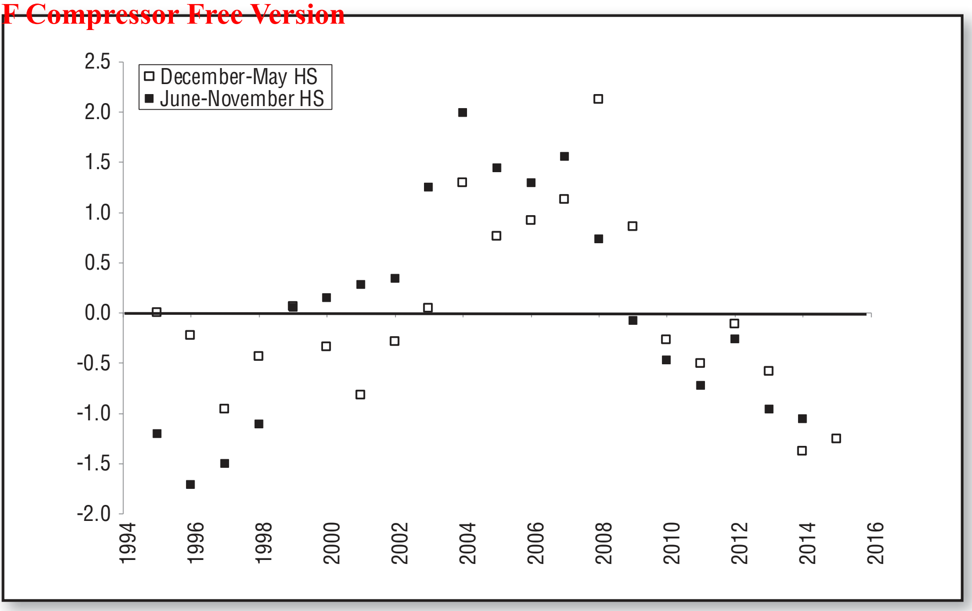


FIGURE E.6 Standardized Residuals for Hog Slaughter vs. Previous Six-Month Pig Crop after Including Dummy Variables

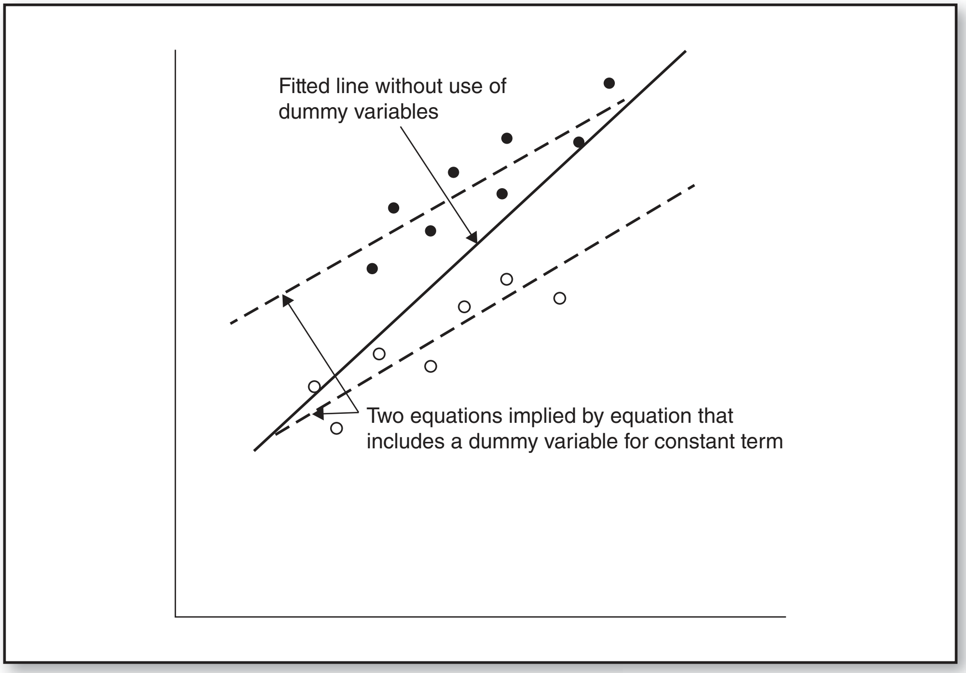

Fitted line without use of dummy variables
Two equations implied by equation that includes a dummy variable for constant term

FIGURE E.7 Bias in Regression Line Due to Omission of Dummy Variables

where S_1 = dummy variable for the first quarter
S_2 = dummy variable for the second quarter
S_3 = dummy variable for the third quarter

Note that the number of dummy variables is always equal to one less than the number of periods, since the base period conditions, assumed to be the fourth quarter in our example, are captured by the original constant and regression coefficient.[4]

If there are two independent variables, the full-version equation would be

$$Y = a + b_1X_1 + b_2X_2 + c_1S_1 + c_2S_2 + c_3S_3 + d_1 \cdot S_1 \cdot X_1 \\ + d_2 \cdot S_1 \cdot X_2 + e_1 \cdot S_2 \cdot X_1 + e_2 \cdot S_2 \cdot X_2 + f_1 \cdot S_3 \cdot X_1 + f_2 \cdot S_3 \cdot X_2$$

Note that:

b values are regression coefficients for regular independent variables.
c values are regression coefficients for dummy constants.
d values are regression coefficients for dummy slope for the first period (S_1).
e values are regression coefficients for dummy slope for the second period (S_2).
f values are regression coefficients for dummy slope for the third period (*S*3).

As should be quite apparent, when the number of periods is increased, the number of dummy variables increases like rabbits. Since the researcher might wish to avoid beginning with an equation that contains a large number of dummy variables, she might prefer to only include constant dummy variable terms in the starting equation and then experiment with the addition of selected slope dummy variable terms if she believes that her initial equation needs improvement.

Multicollinearity

The reader may recall that the extension to the multiple regression model required the additional assumption that the independent variables be linearly independent. *Multicollinearity* is a term used to describe the presence of significant correlation between two or more independent variables.

To see why multicollinearity is a problem, consider a hog-slaughter-forecasting model that includes both the pig crop during the prior six-month period and the number of market hogs at the start of the period as explanatory variables. In this case, the independent variables would be extremely highly correlated, that is, large market hog figures would coincide with large pig-crop numbers. As illustrated in Figure E.8, a three-dimensional representation is really unnecessary for this model, as

[4] In fact, including a dummy variable for the base period would actually result in perfect multicollinearity—a totally undesirable situation (see next section).

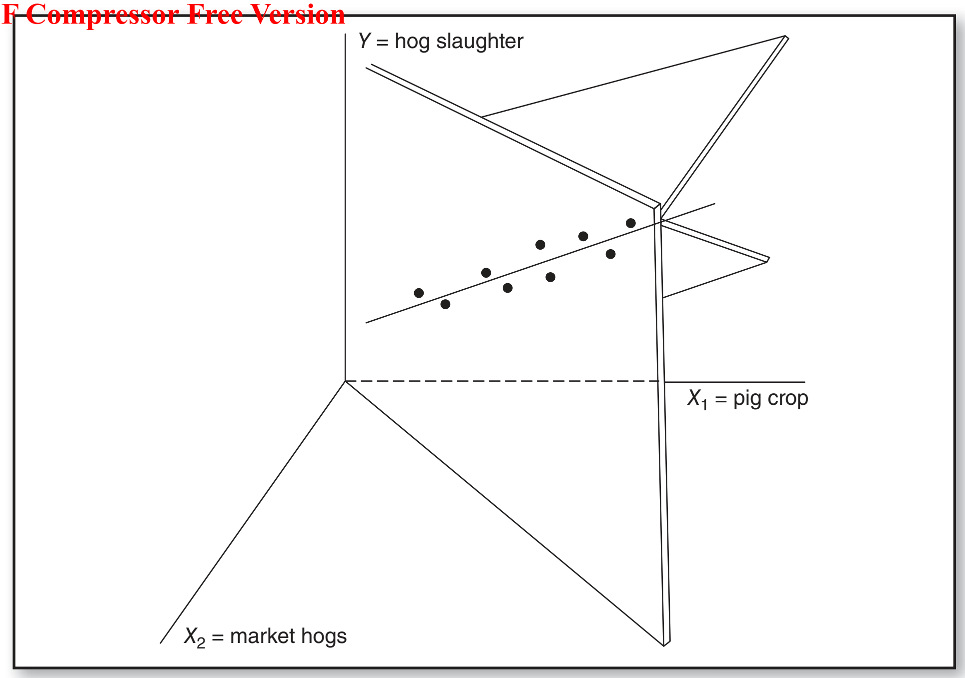


FIGURE E.8 Multicollinearity
Source: Adapted from T. H. Wonnacott and R. J. Wonnacott, *Econometrics*, John Wiley & Sons, New York, 1980.

demonstrated by the proximity of the points to a straight line. Actually, either the $X1$, Y plane or the $X2$, Y plane alone would have been adequate. The first plane would be a two-dimensional representation of the relationship between hog slaughter and the pig crop, and the second, a two-dimensional representation of the relationship between hog slaughter and market hogs. In effect, the inclusion of both the pig crop and market hogs forces the use of a three-dimensional model to represent a relationship that can be adequately described by two dimensions.

The problem lies not in the fact that multicollinearity implies the inclusion of superfluous information, but rather that this redundancy can severely affect the regression equation's reliability. Multicollinearity is a perfect example of the phrase "more is less." As can be seen in Figure E.8, when multicollinearity is present, there may be very divergent planes that closely approximate the fit of points. For any given set of observations, the regression procedure will choose one plane that best fits the observations. However, the real problem in multicollinearity lies in the fact that if the observations were only slightly altered, a totally different plane might be chosen as the best fit. Thus, if multicollinearity exists, the regression coefficients are no longer reliable indicators of how the dependent variable will change when each of the independent variables is changed (while all the other independent variables are held constant). This fact will be reflected by high standard errors, and hence low t statistics, for the regression coefficients of highly correlated explanatory variables.

What about the reliability of the equation in forecasting? If the values of the independent variables for the forecast period lie in the neighborhood of past observations, then the multicollinear model can still provide adequate forecasts. This situation describes the preceding example, since presumably the pig crop and market hogs will continue to remain highly correlated. In other circumstances, however, if the two correlated independent variables cease to be correlated in the future, then the forecast provided by the multicollinear equation could be distorted because the model is only valid for points in the neighborhood of past observations. At other locations, there is no historical evidence to provide any clues regarding the expected relationship between the variables. In geometric terms, all planes passing through a line provide accurate forecasts in the vicinity of the line, but drastically different projections at points removed from the line (Figure E.8).

To summarize, there are two major drawbacks to a multicollinear equation:

1. The regression coefficients lose their meaning (i.e., are no longer statistically reliable).
2. If the equation is used for forecasts in which the independent variables do not lie in the neighborhood of past observations, the projection could be severely distorted.

Clearly, then, it is always desirable to avoid multicollinearity. The presence of multicollinearity can be detected in a number of ways:

1. **Check the regression coefficients.** The regression coefficients of an equation can provide a number of clues indicating that multicollinearity is present:
   a. Low *t* statistics for coefficients that were expected to be highly significant.
   b. In more extreme cases, a regression coefficient sign that may actually be counter to theoretical expectations.
   c. Large changes in the coefficient values when variables are added or deleted from the equation.
   d. Large changes in coefficient values when data points are added or deleted from the equation.

   Any of these patterns would suggest that the independent variables should be examined for signs of correlation.
2. **Compare the independent variables.** Sometimes common sense will dictate when the independent variables are likely to be correlated. By being aware of the problem, one can often avoid multicollinearity by carefully selecting the independent variables. For example, if the researcher thought that gross national product (GNP) and disposable income might help explain the variation of the dependent variable, she would use either one, or try both successively, but she would not include them in the same equation simultaneously. Beyond intuition, one can check for correlation between the independent variables statistically. High absolute values of the *correlation coefficients*[5] between any two independent variables would indicate a potential

[5] The correlation coefficient, denoted by the symbol r, reflects the degree of relationship between two variables and can range between -1 and $+1$. Values close to $+1$ indicate a strong positive relationship, while values close to -1 indicate a strong inverse relationship. If r is close to 0, it means that there is little, if any, correlation between the two variables. The square of the correlation coefficient is equal to the r^2 of the simple regression equation in which one of the variables is a dependent variable and the other an independent variable.

multicollinearity problem. The *correlation matrix*—an output feature in some software packages—offers a summary array of all the paired correlation values.

What should be done if multicollinearity is discovered in an equation? One solution is simply to delete one of the correlated independent variables.

Addendum: Advanced Topics

Transformations to Achieve Linearity[6]

Perhaps the most basic assumption in a regression analysis is that the relationship between the dependent and independent variables is approximately linear. If, in fact, the relationship is decisively nonlinear, the error terms might appear to be correlated. For example, consider what happens when we try to fit a regression line to the scatter points in Figure E.9a. Forcing these points into a linear fit would result in the residual pattern illustrated in Figure E.9b, in which the residuals would tend to be positive at high and low values of the independent variable X and negative in the middle range of values. (In Figure E.9b, the standardized residuals are plotted against the independent variable not time.)

Fortunately, many nonlinear relationships can be transformed into linear equations. For example, the scatter of points in Figure E.9a suggests a hyperbolic function, or an equation of the general form

$$Y = a + \frac{b}{X + c}$$

This can be transformed into a linear relationship by setting

$$X' = \frac{1}{X + c}$$

then

$$Y = a + bX'$$

In this form, the equation can be solved in straightforward fashion using ordinary least squares (OLS), the standard regression procedure. To get a specific forecast for Y, one would merely plug in the value $1/(X + c)$ for X'. For example, if $a = 2$, $b = 16$, $c = 4$, and $X = 4$, the forecast for Y is 4.

[6] Although still involving nothing more mathematically complex than algebra, the remainder of this appendix covers material that is somewhat more advanced.

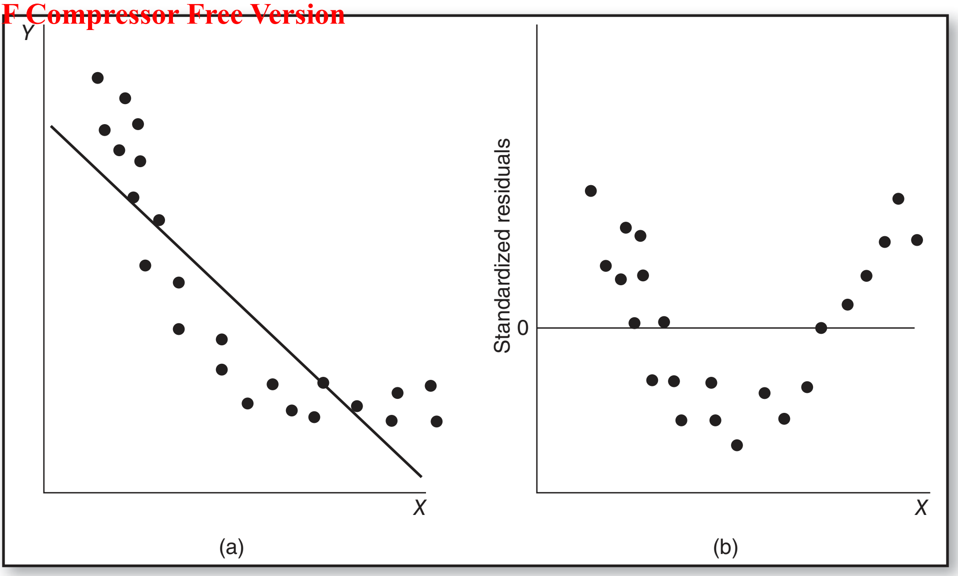


FIGURE E.9 Distortion in Applying Linear Regression to Nonlinear Function

Many other types of functions can be transformed into linear equations. Let us consider a few more examples:

1. $Y = a + b_1X + b_2X^2 + b_3X^3$
   Let $X_1 = X$; $X_2 = X^2$; $X_3 = X^3$; then
   $Y = a + b_1X_1 + b_2X_2 + b_3X_3$
   This is a linear equation and OLS can be applied. Note that although the independent variables are related, the relationship is nonlinear, so that the regression assumption regarding linear independence among the explanatory variables is not violated.
2. $Y = ae^{bX}$
   Taking the natural logarithm of both sides:

$$\ln Y = \ln a + bX$$

Let $Y' = \ln Y$; $a' = \ln a$; then

$$Y' = a' + bX$$

This is a linear equation and OLS can be applied. Note that in this case, plugging a value for X into the derived regression equation will yield a forecast for ln Y. To get a forecast for Y, it would be necessary to find the antilog value. Figure E.10 illustrates the shapes of the function $Y = ae^{bX}$ for different values of b.

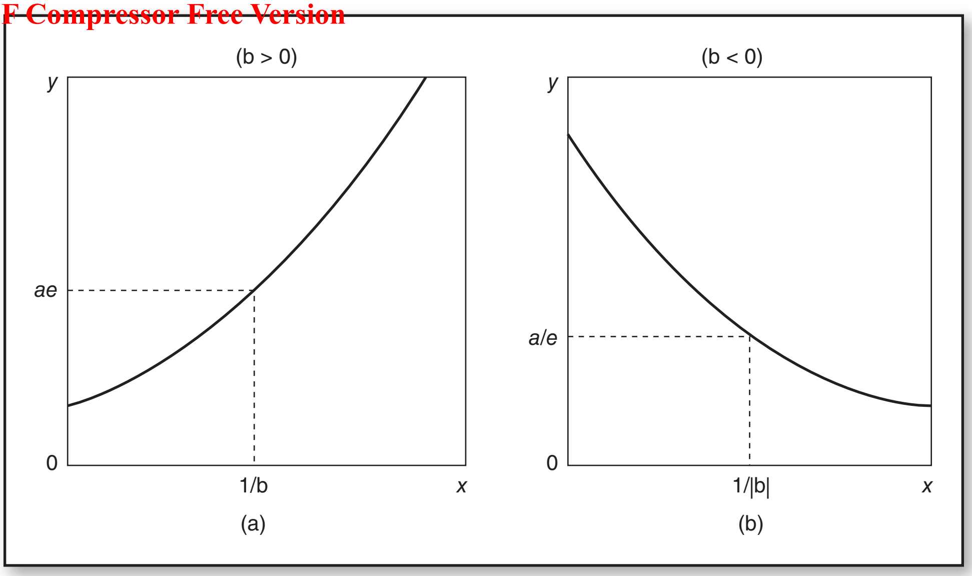


FIGURE E.10 $Y = ae^{bX}$
Source: S. Chatterjee and B. Price, *Regression Analysis by Example,* 3rd ed. (New York, NY: John Wiley & Sons 1999). Copyright © 1999 by John Wiley & Sons; reprinted with permission.

3. $Y = a \cdot X^b$

   Taking logs of both sides:

   $$\log Y = \log a + b \log X$$

   Let $Y' = \log Y$; $a' = \log a$; $X' = \log X$; then

   $$Y' = a' + bX'$$

   This is a linear equation and OLS can be applied. Here we would plug the value for log X, not X, into the regression equation to get a forecast of log Y. To get a forecast for Y, it would then be necessary to find the antilog value. Figure E.11 illustrates the shape of the function $Y = aX^b$ for different values of a and b.

If a residual plot still reflects autocorrelation after all feasible variables have been tried, the possibility of nonlinearity should be considered. In the simple regression case, a scatter diagram can be constructed in order to check whether a linear assumption is warranted or whether another functional form is more appropriate, just as Figure E.9a suggested the equation form $Y = a + b/(X + c)$. In a multiple regression, if nonlinearity is expected for one of the independent variables, a regression could be run using only the other independent variables. The residuals of this equation would then be plotted against the unused independent variable. The presence of any nonlinearity would be apparent in the resulting scatter diagram.

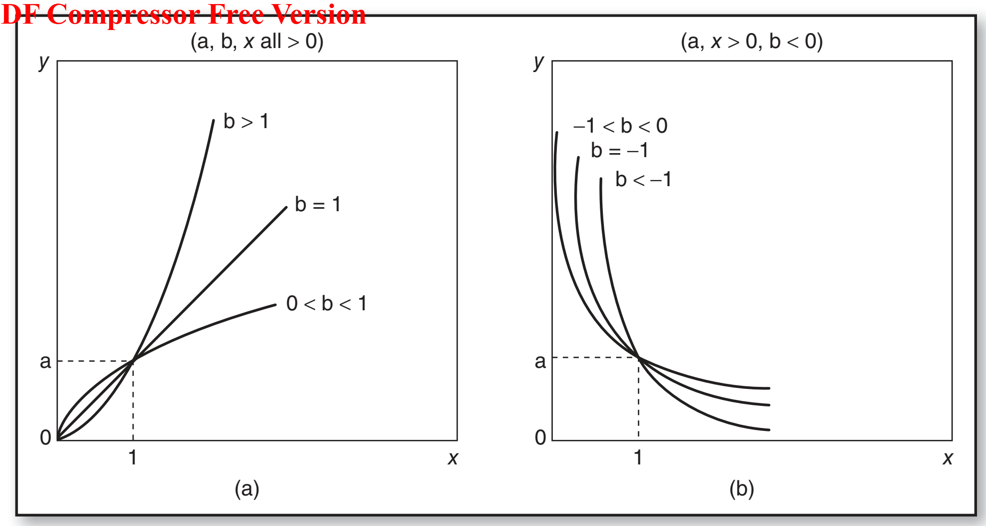


FIGURE E.11 $Y = aX^b$

Source: S. Chatterjee and B. Price, *Regression Analysis by Example*, 3rd ed. (New York, NY: John Wiley & Sons, 1999). Copyright © 1999 by John Wiley & Sons; reprinted with permission.

Transformation to Remove Autocorrelation

The simplest assumption one can make about autocorrelation is that a current period's error term will be equal to the previous period's error term plus a random disturbance. This can be expressed as

$$e_t = e_{t-1} + \upsilon_t,$$

where υ_t = a random disturbance term.

Since $Y_t = \alpha + \beta X_t + e_t$ and $Y_{t-1} = \alpha + \beta X_{t-1} + e_{t-1}$, then

$$Y_t - Y_{t-1} = \beta(X_t - X_{t-1}) + \upsilon_t$$

Let $Y_t^* = Y_t - Y_{t-1}$ and $X_t^* = X_t - X_{t-1}$; then

$$Y_t^* = \beta X_t^* + \upsilon_t$$

For a k-variable multiple regression equation, these steps would yield

$$Y_t^* = \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \cdots + \beta_k X_{kt}^* + \upsilon_t$$

Since by definition υ_t is randomly distributed, OLS can now be applied to this equation.

The preceding method, which is perhaps the most commonly used transformation for removing autocorrelation, is known as *first differences.* In effect, the first difference regression equation states that the change in Y will be linearly dependent on the change in X. Equations of this type will tend to have much lower R^2 values. This is only to be expected, since forecasting the *change* from one period to the next is much more difficult than forecasting the *level.* Once again, consider the following daytrading price-forecasting model:

$$P_t = a + bP_{t-1}$$

where P_t = closing price on day t
P_{t-1} = closing price on day $t - 1$

Such an equation would have an extremely high R^2 since it would give us a close approximation of the price level for a given day. However, it would be useless in forecasting the change in price from day to day. The model

$$P_t^* = a + bX_t^*$$

where $P_t^* = P_t - P_{t-1}$
$X_t^* = X_t - X_{t-1}$

in which X_t is some explanatory variable the value of which is known before day t, would be far preferable, even if its R^2 value were low (e.g., $R^2 = 0.30$).

The first difference approach is easy to use, but it does involve an extreme simplifying assumption regarding the nature of the autocorrelation. A more realistic assumption would be

$$e_t = \rho e_{t-1} + \upsilon_t$$

where $|\rho| < 1$. Note that the larger the value of ρ, the more the error term in a given period will be dependent upon the previous period's error term. A generalized transformation is analogous to the first difference transformation:

$$Y_t = \alpha + \beta X_t + e_t$$
$$Y_{t-1} = \alpha + \beta X_{t-1} + e_{t-1}$$

If we multiply the equation for $Y_t - 1$ by ρ

$$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho e_{t-1}$$

Thus, $Y_t - \rho Y_{t-1} = \alpha(1 - \rho) + \beta(X_t - \rho X_{t-1}) + \upsilon_t$.
Let $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$.
Then $Y_t^* = \alpha(1 - \rho) + \beta X_t^* + \upsilon_t$.

For a k-variable equation, these steps would yield

$$Y^* = \alpha(1-\rho) + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \cdots + \beta_k X_{kt}^* + \upsilon_t$$

Since by definition υ_t is randomly distributed, OLS can once again be used. The only problem with this procedure is that we do not know the value of ρ. We very briefly describe two approaches for estimating ρ.

1. **The Hildreth-Lu procedure.** This procedure specifies a set of spaced values for ρ. If positive autocorrelation is assumed, these values might be 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0. A regression is then run for the transformed equation:

   $$Y_t^* = \alpha(1-\rho) + \beta_1 X_{1t}^* + \beta_2 X_{2t}^* + \cdots + \beta_k X_{kt}^*$$

   using each of the specified values. The procedure will select the equation that results in the lowest SER. If desired, the process can be repeated using a closer spacing of ρ values in the vicinity of the ρ selected in the initial step.
2. **The Cochrane-Orcutt procedure.** This iterative procedure estimates a ρ value from the residuals of the original equation, and a regression is then run on the transformed equation using this estimate of ρ. If the resulting equation still indicates autocorrelation, the process is repeated using the residuals of the new equation.

Heteroscedasticity

One of the assumptions that justifies the use of ordinary least squares (OLS) is that the error terms are homoscedastistic, that is, they have an approximate constant variance. When this condition is not met, the problem is called *heteroscedasticity.* Figure E.12 illustrates a case of heteroscedasticity. Note that the relationship between the dependent and independent variables becomes increasingly variable as X increases, resulting in higher absolute residual values at higher values of X. The wider variability between the dependent and independent variables in a given region will make the resulting regression equation less reliable.

Weighted least squares (*WLS*) is a method used to circumvent this problem. For the relationship depicted in Figure E.12, the WLS approach would give greater weight to the observations for lower values of X, since these offer a more precise indication of the location of the true regression line. Rather than describe the WLS procedure, suffice it to say that there is a simpler approach using a transformation that achieves exactly equivalent results. This transformation assumes that the standard deviation of the error terms is proportional to the independent variable. Specifically,

$$\sigma_i = kX_i$$

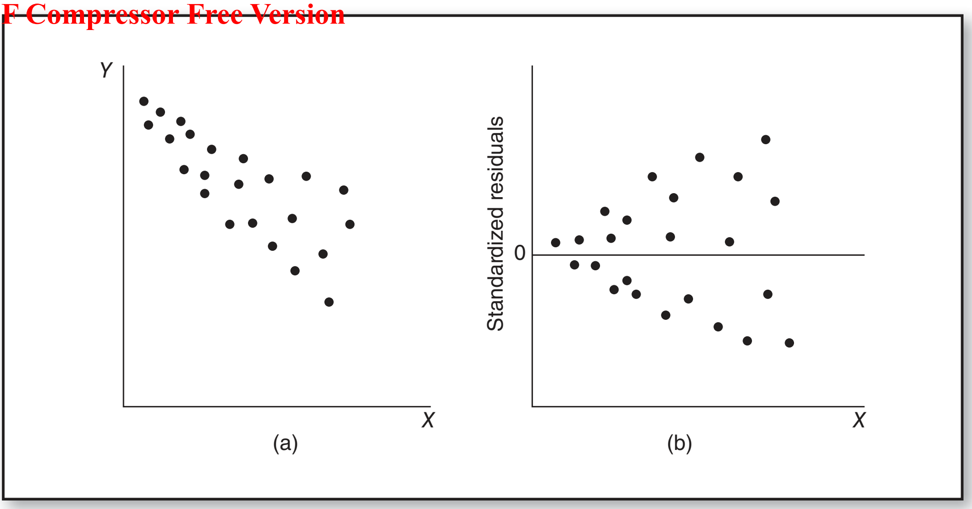


FIGURE E.12 Heteroscedasticity

where σ_i = standard deviation of the error terms (e_i). Starting with the standard regression equation

$$Y_i = \alpha + \beta X_i + e_i$$

we divide by X_i,

$$\frac{Y_i}{X_i} = \frac{a}{X_i} + \beta + \frac{e_i}{X_i}$$

The standard deviation of e_i/X_i. is equal to the standard deviation of e_i divided by X_i. Since the standard deviation of e_i is σ_i, which equals kX_i, the standard deviation of $e_i/X_i = k$, a constant. Thus, this transformation removes the heteroscedasticity of the original equation. Now if we let

$$Y_i' = \frac{Y_i}{X_i} \quad X_i' = \frac{1}{X_i} \quad \alpha' = \beta \quad \beta' = \alpha \quad \text{and} \quad e' = \frac{e_i}{X_i}$$

then

$$Y_i' = \alpha' + \beta' X_i' + e_i'$$

This equation can be solved by using OLS, yielding

$$Y_i' = a + bX_i'$$

where a is an estimator of β in the original equation and b is an estimator of α in the original equation.

APPENDIX F

Practical Considerations in Applying Regression Analysis

I remember the rage I used to feel when a prediction went awry. I could have shouted at the subjects of my experiments, "Behave, damn you, behave as you ought!" Eventually I realized that the subjects were always right. It was I who was wrong. I had made a bad prediction.

—Burrhus Frederic Skinner

Determining the Dependent Variable

The title of this section might sound trivial. After all, the dependent variable is what we wish to forecast. However, in a price-forecasting equation, the selection of a dependent variable is by no means obvious. The following choices must be made:

1. Should the price be stated in nominal or deflated terms?
2. Should the price be based on cash or futures?

3. If the price is based on futures, should it be based on a nearest futures price series or a single contract?
4. Should the price represent the entire season or only a specified segment of the season?

The answer to question 1 would typically be deflated prices, unless a trend variable is included in the equation, in which case nominal prices may be a better choice. If, however, the equation does not include a trend variable, then the use of nominal prices implicitly assumes that equivalent fundamental conditions in two widely spaced years should result in approximately equal price levels. Obviously, this assumption is wrong. All else being equal, inflation will result in considerably higher prices in the more recent year. The subject of adjusting prices for inflation is covered in greater detail in Chapter 25.

The answers to questions 2 and 3 depend primarily on the particular price you wish to forecast. Although this consideration is also a factor in answering question 4, the choice of the time period should depend more heavily on the fundamental characteristics of the market. Of course, if the initial choice is inappropriate, the misjudgment will become apparent in analyzing the regression results. However, by giving some thought to the intrinsic market fundamentals before selecting the forecast period, it is possible to minimize unnecessary trial and error in the regression-analysis process.

For example, in most agricultural markets, the statistical balance for a given season will have a far greater impact on price levels during the first half of the season than on price levels during the second half. This typical market behavioral pattern reflects the fact that by the second half of the season, the prevailing fundamental situation is well-defined and frequently largely discounted. More often than not, major price shifts during the latter part of a season reflect developments affecting new crop expectations (e.g., drought and freeze). Consequently, for a fundamental model that does not include new crop expectations as an explanatory variable, it would generally make more sense to select a price forecast period that approximates the first half of the season rather than the full season. This approach does not mean that we ignore the other six months. Rather, the implication is that it will be necessary to develop other models to forecast prices in those months. For example, the latter months of a season might be grouped with the early months of the following season in a model that employed new-crop statistics to forecast prices.

In some markets, intrinsic fundamental considerations will not dictate a specific observation period. In such cases, the choice will involve only the time frame of the individual observations (e.g., annual, semiannual, quarterly, monthly).[1] Here a general rule might apply: start with the longest period (i.e., annual or semiannual), and if the regression model is satisfactory, work toward a shorter period. Although the shortest time frame projection is most useful for trading purposes, the difficulty of forecasting is inversely proportional to the length of the time period. Also, the shorter the time frame, the more likely the problem of autocorrelation. For example, in a monthly model there is a high probability that a high positive residual in one month will be followed by a positive residual in the next month. Thus, for monthly and even quarterly models, transformations to remove autocorrelation may be necessary (e.g., first differences).

[1] The choice of the length of the period must also be made for markets in which the structure of the model depends on the forecast period. For example, a model based solely on old-crop statistics (i.e., a model that does not incorporate new-crop expectations) could use a six-month forecast period (coinciding with the first half of the season) or it could be applied to two separate three-month periods.

Selecting the Independent Variables

General Considerations

There is more to selecting the independent variables than choosing the factors that intuitively appear to be good candidates for explanatory variables. Perhaps the pivotal question to be considered is whether the regression equation is intended for explaining or forecasting the dependent variable. Sometimes, a regression equation is only intended as an explanatory model. For example, a wheat producer might be interested in determining the relationship between yield and the quantity of fertilizer applied. In this case, his goal is not to forecast yield, a projection that will also depend heavily on other factors, such as weather conditions, but to understand the implications of various management choices. Furthermore, he need not worry about estimating the independent variable (quantity of fertilizer), since it is entirely under his control.

In contrast, most applications of regression analysis in the futures markets will be concerned with forecasting. If an equation is intended primarily for prediction, it is critical to choose explanatory variables that can be determined with relative reliability. For example, if we were to construct a copper price-forecasting model in which the concurrent gross domestic product (GDP) was an important input, the equation would be useless if GDP levels were no more predictable than copper prices, even if $R^2 = 1.00$. Thus, in selecting independent variables, the researcher should keep in mind the precision with which these variables can be estimated *before* the forecast period.

If they prove statistically significant, *lagged variables* are the ideal choice for explanatory variables. A lagged variable is one whose value is determined during a period before the period for the corresponding dependent variable. For example, the average GDP during the prior six months would be a lagged variable. Thus, even if the lagged value of GDP were substantially less significant in explaining copper prices than the concurrent value, it might still be a preferable choice.

Unfortunately, the analyst will rarely be lucky enough to construct a regression equation that uses only lagged variables. Concurrent variables that can be forecast with reasonable accuracy provide an acceptable alternative. In fact, some variables, such as population, can be forecast with such accuracy that they are similar to lagged variables. Other variables can be projected within a reasonable range. For example, in the hog model, hog slaughter is much easier to project than hog prices, since it depends on lagged variables (e.g., prior pig crop, market hogs at start of period). In short, the essential question to consider is whether the values for a potential explanatory variable are known before the forecast period or are at least substantially easier to project than the values for the dependent variable.

Another criterion in selecting the independent variables is that they should not be correlated, in order to avoid the problem of multicollinearity. If several correlated variables seem to be good choices for explanatory variables, they should be tested individually.

Should the Preforecast Period Price Be Included?

An important question to be decided in a price-forecasting equation is whether to include the preforecast period (PFP) price as an explanatory variable. One reason for including the PFP price is that

It is usually an important factor. For example, consider the following two situations in which the PFP price was not taken into account by the model:

Situation A Projected average price for forecast period = 60¢;
price on day before forecast period = 40¢.
Situation B Projected average price for forecast period = 60¢;
price on day before forecast period = 80¢.

Would it be reasonable to expect the same price level in both cases? Definitely not! Some textbook theories notwithstanding, in the real world, prices do not adjust instantaneously to changing fundamentals. In situation A, a major uptrend would be required for prices merely to reach the forecasted equilibrium level. Such an advance will not occur overnight. Furthermore, it is not sufficient for prices to reach 60¢ in order to achieve the projected 60¢ average. Prices would have to go far beyond .60¢ in order to make up for all the days of sub-60¢ prices during the early part of the period. Similarly, in situation B, prices would have to go far below 60¢ to achieve a 60¢ average. In practice, prices may well reach 60¢ in both situations A and B, but the average price is likely to be well below 60¢ in situation A and well above 60¢ in situation B.

The preceding example illustrates that the PFP price may often be an important explanatory variable. Then why not always include it in the model? Ironically, the answer is that it may sometimes be too good in explaining price behavior. In other words, if the PFP price swamps the effect of the other independent variables, the price projection will primarily reflect current price levels. Thus, if the PFP price accounts for a large percentage of the R^2, the model may be good at explaining prices, but will be ineffective at predicting price changes, which after all is the primary goal in price projection. However, in some cases, the other independent variables may explain a major portion of total variation, even when the PFP price is included. In these situations, including the PFP price may help eliminate a significant portion of the unexplained variation that would exist if it were omitted, while still yielding a model that is capable of predicting price changes.

The decision of whether to include the PFP price as an independent variable must be made empirically on a case-by-case basis. A reasonable procedure would be to use a stepwise regression approach (see the section titled "Stepwise Regression" in this appendix) both with and without including the PFP price on the list of independent variables. Although the model that includes the PFP price will always exhibit better summary statistics, it should only be chosen if the effect of the PFP price is significant without being overwhelming.

Choosing the Length of the Survey Period

Ideally, it is desirable to use the longest feasible survey period, since more data points will increase the accuracy of the regression statistics. However, in the real world, there is a tradeoff between the length of the survey period and the relevance of the earliest data points to current conditions. For example, it would be ludicrous to include data before 1973 in a fundamental forecasting model for currency rates, since exchange rates were fixed before that point.

As the preceding example illustrates, fundamental considerations will often limit the number of observations that can be included without distorting the model. Basically, the longest survey period consistent with current market conditions should be used. Scatter diagrams for the dependent variable plotted against each of the explanatory variables may be helpful in making this decision. It will often be necessary to run several regressions for periods of different lengths in order to decide on the optimum number of observations to be included. Occasionally, it may be possible to include earlier nonrepresentative years through the use of dummy variables.

Sources of Forecast Error

In order to build the best model as well as understand its potential limitations, it is important to be aware of the potential sources of forecast error. These include:

1. Random errors for true population regression. Any regression equation is only a simplification that cannot include all possible influences on the dependent variable. Thus, even if we knew the true population regression equation, which we never do, and the explanatory variables were precisely determined, this source of error would still exist. In other words, this type of error can never be avoided.

2. Random errors in the estimated regression coefficients. Since the data used to run a regression represents only a sample from the population, the estimated regression coefficients will deviate from the true population values.

3. Regression equation may be misspecified. The regression model may not represent the true underlying model because of the following reasons:

a. Omission of significant variables;

b. True model is nonlinear or wrong functional form is assumed in a linear transformation;

c. Error terms are autocorrelated[2].

4. Errors in independent variable values. Often, the independent variables must themselves be projected, thereby introducing another tier of potential forecasting error. Sometimes, unexpected events (e.g., droughts, freezes, export embargoes) can result in the actual values of the explanatory variables deviating sharply from the estimated levels. In these situations, the regression projections can prove wide of the mark, even when the model would have provided an accurate forecast if the input had been correct.

5. Data errors. Lagged variable data and the data used to forecast the independent variables may be inaccurate because of sampling or compilation errors.

6. Structural changes. Structural change accounts for perhaps the most serious vulnerability of the regression forecast. Regression analysis is a static approach to a dynamic process; that is, the structure and behavior of a market are constantly changing. Thus, even if a model offers a good representation of the past, it may fail to describe a market adequately in the future. Any major structural change in a market can lead to large forecast errors.

[2] Of course, conditions 3(a) and 3(b) could also result in autocorrelation; the implication here is autocorrelation that exists without the presence of 3(a) and 3(b).

As an example, consider the plight of the unfortunate fundamental analyst using historical regressions to forecast prices for the 1981–1982 period, when the unprecedented combination of severe recession and high interest rates resulted in a dramatic downward shift in demand for many commodities. As a result, prices in a broad spectrum of markets declined to well below the levels that might have been anticipated on the basis of fundamental models that worked well in prior years, but did not include these effects. As a more recent example, the late 2008 financial crisis had such a huge depressant impact on commodity prices across the board that virtually any viable fundamental model for any commodity market would have been likely to yield price forecasts for the late 2008, early 2009 period that were far too high.

The preceding examples illustrated structural changes simultaneously affecting a broad range of markets. A structural change can also be confined to a single market. One example of such a change was the dramatic shift in corn usage for ethanol production. Corn use for fuel went from one-tenth the feed-use level before 2000 to greater than feed usage by 2010.

It is important to realize the standard error measures in regression analysis only account for the first two sources of error just listed. Perhaps even more sobering is the fact that with the exception of a misspecified equation (3), all these sources or error are beyond the control of the analyst. However, the potential variability attributable to errors in estimating the independent variables (4) can at least be defined by allowing for a range of possibilities. For example, in addition to generating a price forecast based on a set of best estimates for the explanatory variables, projections can also be derived for sets of bearish and bullish assumptions. In this way, it is at least possible to gauge the potential impact of inaccurate estimates for the independent variables. Furthermore, some solace can be drawn from the fact that the various types of errors listed here are not necessarily cumulative; that is, they may partly offset each other.

As a final word, it should be emphasized that this list of potential errors is not intended to discourage the potential user of regression analysis, but rather to instill a sense of realism in interpreting regression results.

Simulation

As the previous section demonstrated, comparisons between the fitted values of the regression equation and actual observations may severely understate potential forecasting errors. The process of determining how forecasts based on the given model would have compared with reality is called *simulation,* which is an extremely useful technique for testing a model under near real-life conditions. Simulation should only be undertaken once the choice of a model has been finalized, or at least reduced to a small number of possibilities. Ideally, the simulation period should be long enough to include a variety of conditions (e.g., at least one bull, one bear, and one neutral market in a price-forecasting equation).

For example, assume it is 2015 and we have decided the past 20 years of data are relevant to the current market structure. Given the constraint that each forecast must be based on at least 10 years of data, a 10-year simulation of a calendar-year forecast could be constructed as follows:

1. Using only data available on January 1, 2005, derive a regression equation for the same model for 1995 through 2004.

2. Using only data available on January 1, 2005, estimate the independent variables.
3. Plug these values into the 1995–2004 regression equation to obtain a forecast for 2005.
4. Repeat an analogous procedure for each subsequent year (2006–2014).
5. Compare simulations to actual values and calculate the root mean square (defined later in this section).

For a quarterly model, the simulation procedure would be analogous. However, with a quarterly model, very little would be lost by revising the regression equation only once every four times (each year) in order to reduce the amount of computation.

It may be instructive to compare the differences between the simulation forecasts and actual values with the residuals of the current regression equation. Of course, the former will almost invariably be higher, since simulation results are based on forecasts, while the regression equation is a best fit of past values.

A measure that may be useful in comparing the simulation results of different models is the *root mean square* (RMS):

$$\text{rms} = \sqrt{\frac{\sum_{t=1}^{N}\left(Y_t^F - Y_t^A\right)^2}{N}}$$

where

Y_t^F = forecasted value of Y for period t
Y_t^A = actual value of Y for period t
N = number of simulated observations

Note that the RMS calculation is analogous to the formula for the standard error of the regression (SER) (except for the number of degrees of freedom) and reflects the same underlying meaning.

Stepwise Regression

Ideally, having selected a list of explanatory variables, regression equations would be generated for each possible equation form. For example, given a dependent variable Y and three independent variables X_1, X_2, and X_3 there would be eight possible equations:

1. Y vs. X_1, X_2, and X_3 (all independent variables included)
2. Y vs. X_1, X_2
3. Y vs. X_1, X_3
4. Y vs. X_2, X_3
5. Y vs. X_1
6. Y vs. X_2
7. Y vs. X_3
8. $Y = \overline{Y}$ (no independent variables included)

Such a procedure, however, is not very efficient. The total number of possible equations doubles with the addition of each independent variable (e.g., 16 for four variables, 32 for five).

Stepwise regression is a highly useful and efficient procedure for isolating and providing summary results for the most statistically interesting equations. There are two basic types of stepwise regression:

1. **Forward selection.** The program selects the single independent variable that provides the highest r^2 value to form the first equation. Explanatory variables are then added one at a time to form subsequent equations, with the choice depending on which variable will result in the highest R^2 equation. The program terminates with an equation that includes all of the specified explanatory variables.
2. **Backward elimination.** The program begins by listing the equation that includes all the specified independent variables. The program then deletes the variable with the lowest t value to form the second equation. Subsequent equations are formed by continuing to delete variables, one at a time, with the elimination decision dependent on which remaining variable has the lowest t value.

The two methods will not necessarily yield the same subset of equations. Overall, the backward elimination process is preferable, particularly if the PFP price is one of the explanatory variables. In the forward selection process, the PFP price will usually be chosen first, since it is likely to explain more variation in the dependent variable than any other single variable. However, once more explanatory variables are added, the significance of the PFP price may drop sharply, as other variables in combination explain a portion of the variation originally attributed to the PFP price. Thus, in the backward elimination process, at some stage the PFP price might have a lower t value than any of the remaining variables.

Although the PFP price is effective as an explanatory variable, its inclusion may yield equations that are less useful for forecasting purposes. With the forward selection process, there is a higher probability that all of the chosen equations will include the PFP price, since the first variable chosen remains in all subsequent equations.

Once the stepwise regression results have been analyzed, detail should be generated for the equations that appear to be the most promising.[3] Minimum detail would include a listing of actual observations, predicted values, and residuals. Residual plots should also be constructed for these equations and modification implemented as suggested by these plots.

Sample Step-by-Step Regression Procedure

There is no single right order in which to perform the various elements of regression analysis. The following order merely represents one suggested sequence:

1. Determine the dependent variable.
2. List all possible choices for explanatory variables.

[3] The summary statistics would not be the only criteria for making this choice. For example, an equation that did not include the PFP price as a dependent variable might be preferable to one that did if the summary statistics were only modestly less favorable.

3. Choose a subset of these (usually no more than five), taking care to avoid selecting correlated independent variables. Scatter diagrams can be used as an aid in this selection process.
4. Choose the length of the survey period. Scatter diagrams can also be used as an aid in this step.
5. Apply a stepwise regression program to the selected variables.
6. Analyze the results by examining the various key statistics: t values, SER, CR^2, F, and DW. If there is any evidence of multicollinearity, check out this possibility and rerun stepwise regression with a different set of variables if necessary.
7. Generate detail and construct residual plots for the most promising equations in the stepwise regression run.
8. Check residual plots for outliers. Decide whether outliers should be deleted.
9. Check residual plots for autocorrelation.
10. If outliers or autocorrelation exist, try to correct through the addition of variables or transformations to achieve linearity.
11. If autocorrelation is still a problem, try a transformation to eliminate autocorrelation (e.g., first differences).
12. Check the correlation matrix or R^2 values for various combinations of equations based on the explanatory variables in order to verify that multicollinearity is not a problem.
13. Repeat steps 3–12 for other selections of explanatory variables.
14. *Optional:* After narrowing the number of possible models to three or less, generate simulations.

Summary

Regression analysis is an extremely efficient and powerful tool; it is a virtual necessity for fundamental analysis. The foregoing appendices were intended to provide the necessary background to interpret and analyze the results available on standard regression software packages. Regression analysis provides the means for precisely answering the question: What is the approximate equilibrium level, *given the specified conditions and assumptions*? The italicized qualification is essential. There is a danger of viewing regression results with great rigidity because of the scientific manner in which they are derived. This would be a mistake. As explained in the section "Sources of Forecast Error," a variety of factors are capable of causing the regression projection to be inaccurate. Therefore, the trader must always be open to the possibility the regression forecast might be wrong. However, given such a sense of realistic awareness, fundamental regression models can provide valuable insight into a market's current state and its potential future direction.

REFERENCES AND RECOMMENDED READINGS

Wonnacott, R. J., and T. H. Wonnacott. *Econometrics* (New York, NY: John Wiley & Sons, 1980). This is an extraordinarily lucid treatment of an abstruse subject and is an excellent choice for readers interested in a more in-depth understanding of regression analysis. One of the outstanding features of this book is that it is divided into two separate parts, which cover essentially the same material but on different levels of difficulty. As a result, Part I, which provides a comprehensive and insightful overview of the key concepts of regression analysis, is fully accessible to a reader with only limited mathematical background.

Chatterjee, Samprit, and Ali S. Hadi. *Regression Analysis by Example,* 5th edition (New Delhi: Wiley India, 2012). This may be the best book available on the practical application of regression analysis. As promised in the title, the essential concepts are demonstrated by example. Perhaps the book's best feature is its thorough exposition of the use and interpretation of residual plots, an extremely effective yet easy-to-apply method for analyzing regression results.

Pindyck, R. S., and D. L. Rubinfeld. *Econometric Models and Econometric Forecasts,* 4th edition (New York, NY: McGraw-Hill/Irwin, 1997). The first of the three sections in this book covers single-equation regression analysis. (The other two sections are Multi-Equation Simulation Models and Time Series Models.) This book offers a clear exposition of theoretical concepts, as well as many useful insights into the practical application of regression analysis. Readers with limited mathematical background will find the presentation more difficult than Part I of Wonnacott and Wonnacott.

Makridakis, S., and S. C. Wheelwright. *Forecasting Methods and Applications,* 3rd edition (New York, NY: John Wiley & Sons, 1997). This text provides a broad overview of forecasting techniques, with regression analysis accounting for one of six sections. The presentation is aimed at an audience interested in practical applications rather than theory. This book is clearly written, covers a wide range of topics, and provides a plethora of examples to illustrate the discussion.

Freund, J. E., and F. J. Williams. *Elementary Business Statistics: The Modern Approach*, 6th sub edition (Upper Saddle River, NJ: Prentice Hall College Div., 1992). This book provides a good general overview of elementary statistics for the nonmathematical reader. The text is clearly written and replete with examples.

Kimble, G. A. *How to Use (and Misuse) Statistics* (Englewood Cliffs, NJ: Prentice-Hall, 1978). This introduction to elementary statistics is written with style and a sense of humor. Although it may be hard to believe, this is one statistics book that can actually be read for entertainment value alone.

INDEX

WILEY END USER LICENSE AGREEMENT