DARDEN
BUSINESS PUBLISHING
UNIVERSITY of VIRGINIA

REGRESSION: FORECASTING USING EXPLANATORY FACTORS

Jerry Baugher, sales manager for Lawrence Construction, had just finished gathering data on the company's 30 most recent single-family homes (see **Table 1**). Early in conversations with perspective customers, Baugher was often asked about the company's typical cost per square foot. Baugher always responded with a list of factors affecting the final cost of a new house and a caution against relying too heavily on square footage as the sole determinate of cost. Having said that, however, he knew customers would be pressing him for some numbers.

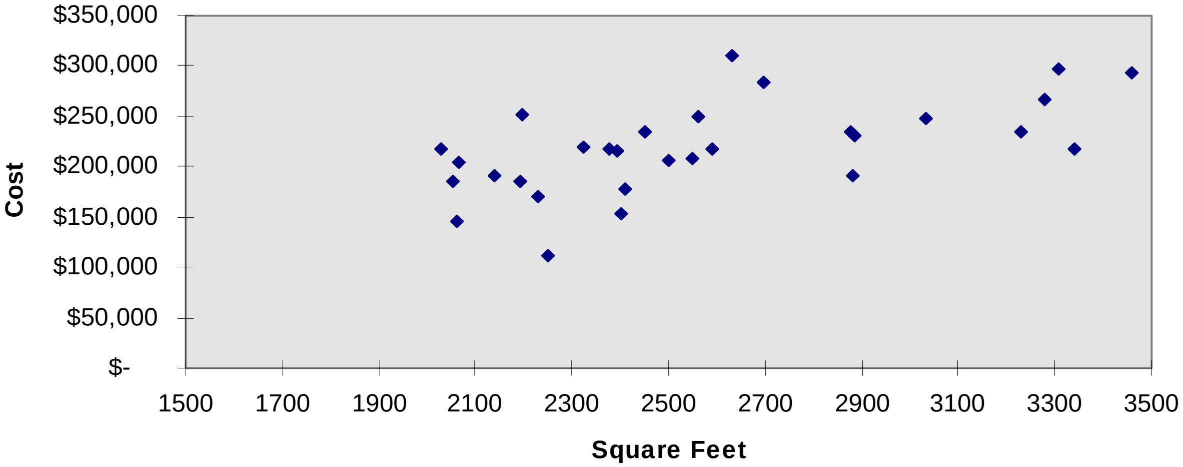
Table 1.
Lawrence Construction Cost Data.

Sq. Foot	Cost	Sq. Foot	Cost
2,885	$231,000	2,323	$219,000
2,230	$170,000	3,279	$267,000
2,377	$217,000	3,308	$297,000
2,551	$209,000	2,140	$192,000
2,592	$218,000	2,029	$217,000
3,341	$218,000	2,879	$234,000
2,632	$310,000	2,408	$177,000
3,034	$248,000	3,228	$234,000
2,052	$186,000	2,560	$249,000
2,066	$204,000	2,500	$206,000
2,193	$185,000	2,248	$112,000
2,392	$216,000	2,696	$283,000
2,060	$146,000	2,400	$153,000
3,457	$293,000	2,196	$251,000
2,450	$234,000	2,880	$191,000
Average		2579.5	$218,900
St. Dev.		429.56	$45,256

A scatter plot of cost versus square feet (**Figure 1**) shows that there was indeed a positive relationship between these two variables. The bigger the home, the more it cost. In keeping with Baugher's understanding of the components of cost, however, the plot suggested to him that smaller homes cost more per square foot than larger homes. This reflected the fact that many components of a home's construction costs had nothing to do with the size of the home. Consequently, a single average cost per square foot such as $84.50 (calculated as $218,900, the average cost of the 30 homes, divided by 2,579.5, the average square feet per home) would understate the cost of a small home and overstate the cost of a large home.

The material comprising this note can also be found in chapter 14 of *Quantitative Analysis: Text and Cases*, Sam Bodily, Robert Carraway, Sherwood Frey, and Phillip Pfeifer, McGraw Hill, 2002. The material in this note is substantially similar to the material in notes QA-0600, QA-0271, QA-0377, QA-0329, and QA-0330. *To order copies, send an e-mail to* sales@dardenbusinesspublishing.com. *No part of this publication may be reproduced, stored in a retrieval system, used in a spreadsheet, or transmitted in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the permission of the Darden School Foundation.* Rev. 10/07.

Figure 1. Scatter plot of cost versus area in square feet.



1. The simple linear model

The situation in Lawrence Construction is typical of many business forecasting problems. A variable of interest, call it *Y*, is known to be related to a second variable, *X*. The decision-maker knows the history of observed *X* and *Y* pairs and wishes to use the known value of *X* in the current situation to forecast the uncertain quantity *Y*. Because the forecast of *Y* will depend on *X*, *Y* is called the *dependent variable* and *X* is called the *independent variable*.

The first step in solving the forecasting problem is to hypothesize the form of the relationship between *Y* and *X*. We need a simple, logical equation that describes how *Y* is related to *X*. Such an equation is called a model.

One simple and often-used model assumes that the underlying relationship between *Y* and *X* is linear. In other words, the equation that shows the relationship of *Y* to *X* is the equation for graphing a straight line. Furthermore, the model should recognize that this relationship is not perfect. The scatter plot of *Y* versus *X* is almost never a perfectly straight line with all points falling on the line. The scatter plot is typically like that in **Figure 1**, with points scattered about some imaginary straight line.

The model we have just described is called the *simple linear model*. In equation form we write it as

$$Y = a + bX + error. \qquad (1)$$

Notice that this equation says that *Y* is equal to *a*, a constant amount, plus the product of *b* times *X*, plus an error. This equation says that *Y* is linearly related to *X*, but that the relationship is not perfect.[1] The error term in the model reminds us that in our uncertain world, Y will not exactly equal $a + bX$.

Let us focus, for a moment, on the $a + bX$ component of our model. This $a + bX$ component is the equation for the imaginary line, which we hypothesize describes the underlying relationship between Y and X. Besides "imaginary line" and "underlying relationship," is there some other way to label just what $a + bX$ represents? Yes. The $a + bX$ component represents the mean of Y for a given X:

$$\text{Mean of } Y \mid X = a + bX. \qquad (2)$$

Recall from sampling theory that the mean is the long-run average of some population. In sampling, we usually talked about a single population and a single mean. Now in equation (2) we shift to a situation where, rather than a single mean, the mean changes depending on the value of X.[2] The imaginary line that describes the underlying relationship between Y and X is a line of means.

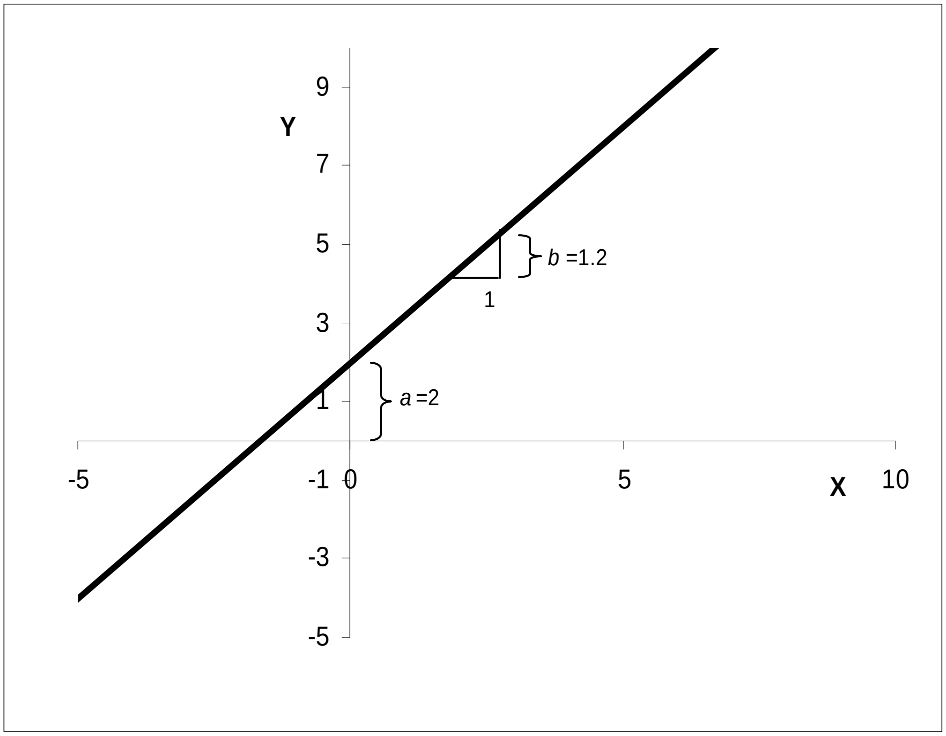
Equations (1) and (2) are two ways to say the same thing. Equation (1) is a model for Y. Equation (2) is a model for the mean of Y given X. Both equations assume that Y is related to X. Equation (2) says the mean of Y is linearly related to X, and equation (1) simply goes one step further and says that Y is equal to its mean plus some error.

The constants a and b are called *coefficients*. They determine just exactly what the linear relationship between Y and X looks like. For example, if $a = 2$ and $b = 1.2$, the graph of the mean of Y versus X is as shown in **Figure 2**. The a coefficient is called the *intercept*, because it is the value at which the line intercepts the Y axis. It is the value of the mean of Y when $X = 0$. The b coefficient is called the *slope* because it measures the pitch or slope of the line. That is, b is the amount by which the mean of Y changes if X increases one unit. If b is positive, the mean of Y increases as X increases and the line goes from the lower left of the graph to the upper right. Conversely, if b is negative, the mean of Y decreases as X increases and the line goes from upper left to lower right. If b happens to equal zero, then the line is horizontal and the mean of Y does not change as X increases. In this case, X is of no use in forecasting Y.

[1] We chose to use $Y = a + bX$ as the equation of a line. This is the convention used by most statistical texts and statistical software packages. From algebra, you may remember the equation of a line as $Y = mX + b$. Rest assured, both are correct. In fact, they are equivalent. We use a as the intercept and b as the slope. In the $Y = mX + b$ expression, b is the intercept and m is the slope.

[2] For those of you who like to think in terms of urns, in sampling we had one urn and one mean. Now we have a whole series of urns, one for each possible value of X. The mean of each urn in this series of urns is given by equation (2).

Figure 2. Mean of $Y|X = 2 + 1.2X$.



2. Fitting the model using "least squares"

After the model form has been chosen, the next step is to determine the particular coefficient values, *a* and *b*, that are best suited for the current situation. A past history of relevant *X*, *Y* pairs is used to guide this selection, which is called fitting the model to a set of data. Fitting the straight-line model means determining exactly how to draw a straight line through the scatter of *Y* versus *X* points (**Figure 1**, for example). The process of using data to "fit the line" or "estimate the coefficients of the model" is sometimes referred to as "regressing *Y* on *X*."

Since the ultimate objective of this modeling is to forecast Y, a good way to fit the model is to pick coefficients that would have done the best job of forecasting past actuals. An accepted measure of forecasting accuracy is *mean squared error*, the average squared difference between actuals and forecasts. The averaging is done on squared differences so that positive and negative errors do not cancel each other out and so that the relative severity of large errors is emphasized.[3]

This same criterion is used to fit models to past data. The procedure known as *least squares* chooses coefficient values that minimize the sum (or average) squared differences between actuals and fitted values. Graphically, *least squares* chooses the fitted line that minimizes the sum of the squared vertical distances between each point (actual) and the line

[3] Statistical theory also dictates mean squared error as the best measure of forecasting accuracy if the forecast errors are unbiased and normally distributed.

(fitted value). These differences between actuals and fitted values are called *residuals*. The resulting line that minimizes the sum of squared residuals is called the *least-squares regression line*.[4]

Table 2. Excel Summary Regression Output.

Regression Statistics	
Multiple R	0.599
R Square	0.359
Adjusted R Square	0.3360
Standard Error	36878.6
Observations	30

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	2.1314E+10	2.13E+10	15.67
Residual	28	3.8081E+10	1.36E+09	
Total	29	5.9395E+10		

	Coefficients	*Std. Error*	*t-stat*	*p-value*
Intercept	56103.66	41670.92	1.35	0.1890
Sq Foot	63.11	15.94	3.96	0.0005

The computations necessary to compute least-squares coefficient estimates are very straightforward. Most electronic-spreadsheet software packages can carry out the necessary calculations to determine the least-squares regression line.

Table 2 shows the results from using Excel's regression feature with the data in **Table 1**. Amid the wealth of numbers in this report (many of which, believe it or not, you will come to appreciate) are the two we want. The estimated intercept, $\hat{a}$, is $56,104. The estimated slope, $\hat{b}$, is 63.11.[5] Thus the least-squares regression line relating Lawrence's new-home construction cost to home size measured in square feet is

$$\hat{Y} = \$56{,}104 + 63.11(X) \qquad (3)$$

where $\hat{Y}$ is the proposed forecast of cost if X is the area of the house in square feet.

Figure 3 shows this line superimposed on the scatter plot of the data in **Table 1**. This particular line is the best line we can draw through the 30 points—best in the sense that the 30 points are "closer" to this line than any other possible line. Our measure of closeness is the sum (average) of squared distances of the points from the line.

[4] The term "regression" comes from the pioneering work of Sir Francis Galton (1822–1911) who studied the strength of the resemblance between parents and their offspring. The best fitting line through a scatter plot of heights of sons versus heights of their fathers revealed that sons' heights regressed toward the overall mean height. Tall fathers tended to have shorter sons and short fathers tended to have taller sons. Because of this phenomenon, the best fitting line was labeled the regression line. The name stuck despite the fact that the regression-to-the-mean phenomenon was a characteristic of the situation being studied rather than the technique used for finding the best fitting line.

[5] The terms $\hat{a}$ and $\hat{b}$ are used to refer to specific estimates of a and b in model (1).

Figure 3. Chart of regression line.

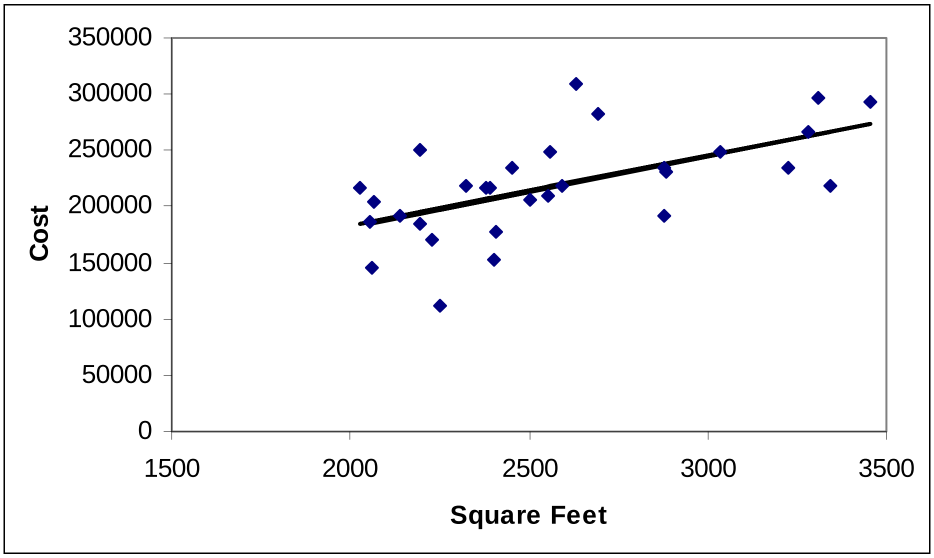


Table 3 presents the regression results in table form. The first two columns contain the original data. The third column contains the fitted values, $\hat{Y}$, found by substituting each X into the regression equation. The last column contains the residuals—the difference between Y and $\hat{Y}$. These residuals are the "distances" of the points from the line. They measure how close the fitted values (the $\hat{Y}$ s) come to the actual values (the *Y*s). Remember that the sum of squares of these residuals was the criterion we used to select $\hat{a}$ and $\hat{b}$.

Now that we have the least-squares regression line, we can use it to provide a point forecast for construction cost of any new home. For example, to estimate the cost of a new home with an area of 2,500 square feet, we substitute X = 2,500 into the least-squares regression equation (3) to obtain $\hat{Y}$ = $213,879. Thus our best guess for construction cost for this new home is $213,879. Additionally, the regression results suggest to Baugher that a simple cost-per-square-foot forecast will not be as accurate as the new forecasting rule of thumb suggested by model 3. A better rule of thumb would use $56,000 (rounding off) plus $63 per square foot to estimate the cost of a Lawrence Construction new home. Now Baugher must decide whether, and how, to use this new forecast in conversations with perspective buyers.

3. Important properties of the least-squares regression line

In **Table 3**, we see that the average of the residuals is zero. Stop and think about what that means. Since residuals measure the differences between the past actuals and fitted values, an average of zero for the residuals means that the fitted values were, on average, neither too high nor too low. On average, the fitted values were "right on." Because the average of the residuals is zero, we don't have to worry about correcting the fitted values (and future forecasts) for bias.

Table 3.
Fitted and Residual Values.

	Sq. Ft.	**Cost**	**Fitted**	**Residuals**
	X	Y	$\hat{Y}$	$Y - \hat{Y}$
	2,885	$231,000	$238,178.2	-$7,178.2
	2,230	$170,000	$196,840.7	-$26,840.7
	2,377	$217,000	$206,118.0	$10,882.0
	2,551	$209,000	$217,099.2	-$8,099.2
	2,592	$218,000	$219,686.8	-$1,686.8
	3,341	$218,000	$266,956.7	-$48,956.7
	2,632	$310,000	$222,211.2	$87,788.8
	3,034	$248,000	$247,581.7	$418.3
	2,052	$186,000	$185,607.0	$393.0
	2,066	$204,000	$186,490.5	$17,509.5
	2,193	$185,000	$194,505.6	-$9,505.6
	2,392	$216,000	$207,064.6	$8,935.4
	2,060	$146,000	$186,111.9	-$40,111.9
	3,457	$293,000	$274,277.6	$18,722.4
	2,450	$234,000	$210,725.1	$23,274.9
	2,323	$219,000	$202,710.0	$16,290.0
	3,279	$267,000	$263,043.9	$3,956.1
	3,308	$297,000	$264,874.1	$32,125.9
	2,140	$192,000	$191,160.7	$839.3
	2,029	$217,000	$184,155.4	$32,844.6
	2,879	$234,000	$237,799.6	-$3,799.6
	2,408	$177,000	$208,074.4	-$31,074.4
	3,228	$234,000	$259,825.2	-$25,825.2
	2,560	$249,000	$217,667.2	$31,332.8
	2,500	$206,000	$213,880.6	-$7,880.6
	2,248	$112,000	$197,976.7	-$85,976.7
	2,696	$283,000	$226,250.3	$56,749.7
	2,400	$153,000	$207,569.5	-$54,569.5
	2,196	$251,000	$194,694.9	$56,305.1
	2,880	$191,000	$237,862.7	-$46,862.7
Average	2,579.5	$218,900.0	$218,900.0	$0.0
St. Dev.	429.6	$45,255.9	$27,110.1	$36,237.2

Were we just lucky, or will this happen all the time? It turns out that this desirable property is, indeed, shared by all least-squares regression lines.

Residuals from a least-squares regression line sum (average) to zero.

In an effort to minimize the sum of squared residuals, the least-squares technique naturally selects $\hat{a}$ to ensure that the residuals sum to zero.

A related property of the least-squares regression line is that it always goes through the point $\overline{X}$,$\overline{Y}$.[6] Another way to say the same thing is

The least-squares forecast of Y at $X = \overline{X}$ is $\overline{Y}$.

Verify for yourself that substituting $\overline{X}$ = 2,579.53 into equation (3) produces a forecast equal to \$218.900, the average construction cost. Some people use this property to say "the regression line goes through the center of the data." It is *not* a coincidence then, that the average of the fitted values in **Table 3** (\$218.900) is equal to the average construction cost of the 30 homes.

4. Summary regression statistics

Put yourself back in the shoes of Jerry Baugher. You carefully compiled the data in **Table 1** in an attempt to understand how Lawrence's construction costs vary with home size. The chart in **Figure 1** confirmed your belief that construction costs were highly dependent on home size. The chart also helped you judge that the underlying relationship was linear and led you to fit the model summarized by equation (1). You scanned the output in **Table 2** for the numbers you needed to forecast cost ($\hat{a}$ = \$56,104 and $\hat{b}$ = 63.11). The large positive intercept confirmed your belief that many cost components of a new home were constant and did not increase with the size of the home. For good measure, you examined the regression line both in graphical (**Figure 3**) and tabular (**Table 3**) form.

Now what? Are we done? How well did we do? How good is this model? Have we made any mistakes? What are all those extra numbers in **Table 2** for? Do we have to know what they all mean?

First of all, much of the information in **Table 2** is redundant. You really don't have to look at all of it. Secondly, this note will attempt to explain the more important summary statistics you find in **Table 2**.

[6] The terms $\overline{X}$ and $\overline{Y}$ are used to refer to the sample averages of X and Y, respectively.

Standard error

After reading the values of the estimated coefficients, we suggest you next look at the standard error.[7] In the Lawrence Construction example, the standard error is 36,879. Just what does this number tell us?

Recall from equation (1) that we recognized that the *Y*s would not fall exactly on the imaginary line. The differences between *Y* and the imaginary line are called errors. The standard error is our estimate, based on these data, of the standard deviation of the errors. The standard error is a measure of how close the points are to the imaginary line. If the errors are normally distributed, we would say that 68% of the errors would be within plus or minus 36,879 of the imaginary line.

Because the standard error measures how close the points are to the imaginary line, it is a measure of how well the model fit the data. The lower the standard error, the closer the points are to the imaginary line, and the better the fit.

The standard error is an *absolute* measure. We should interpret the 36,879 as 36,879 *dollars*. The standard error is always measured in the same units as *Y*. To judge how big this standard error is, you should compare it to the magnitude of Lawrence Construction's costs, which ran from $112,000 to $310,000.

One final interpretation of the standard error has to do with forecasting. Suppose in addition to the point forecast, Baugher required a probabilistic forecast. How much uncertainty is there in the construction cost of a new home? What standard deviation should we use when we forecast construction cost? The simple answer is: Use the standard error, $36,879.[8] Thus the standard error is a measure of how much uncertainty we face when using the model to forecast *Y*.

Adjusted R square

Whereas the standard error is an *absolute* measure of how well the model fits the data, adjusted R square is a *relative* measure. To judge whether the standard error of $36,879 is big or small, you need to know something about the Y variable and the surrounding context.

[7] This statistics has a variety of names. It is sometimes called the *standard error of the Y-estimate*, the *standard error*, or the *standard deviation of residuals*.

[8] This is the simple answer, but not the totally correct answer. Since the standard error measures the standard deviation of the errors in model (1), we can use it as the standard deviation when we forecast only if we know the true imaginary line. If our sample size was huge (and a sample size of 20 is not huge), we could safely assume that $\hat{a}$ and $\hat{b}$ were essentially identical to the true *a* and *b*. With a huge sample size, we would "know" the true line and would use the standard error as the standard deviation when we forecast. When the sample size is not huge, the sta ndard error (although still our best guess for the standard deviation of the errors in the model) understates the standard deviation when we use $\hat{Y}$ to forecast the next *Y*. Why? Because $\hat{Y}$ is not the same as the mean of $Y|X$. Why? Because $\hat{a}$ and $\hat{b}$ are not the same as *a* and *b*. We use an estimated line when we forecast. And the standard error applies to errors about the true, imaginary line.

To put the \$36,879 in context we need a benchmark. If \$36,879 measures how well the model did, we need to measure how well we would do without the model. If we did not use the model to forecast, what would we do? Without the model, we might resort to using $\overline{Y}$ = \$218,900 as the forecast of cost. Remember that the model is how we incorporated information about the area of the home. If we are not to use this information, we have no choice but to treat all homes the same. Our forecast would be \$218,900, and it would not change based on the size of the home. If $\overline{Y}$ is our forecast, the standard deviation of the *Y*s is the uncertainty relevant to our forecast. With Lawrence Construction, this number is \$45,256 (reported in **Table 1**).

We are now in a position to compare the \$36,879 (a measure of how well the model does) to the \$45,256 (a measure of how well we can do without the model). Adjusted R square is the accepted way to compare these two numbers:

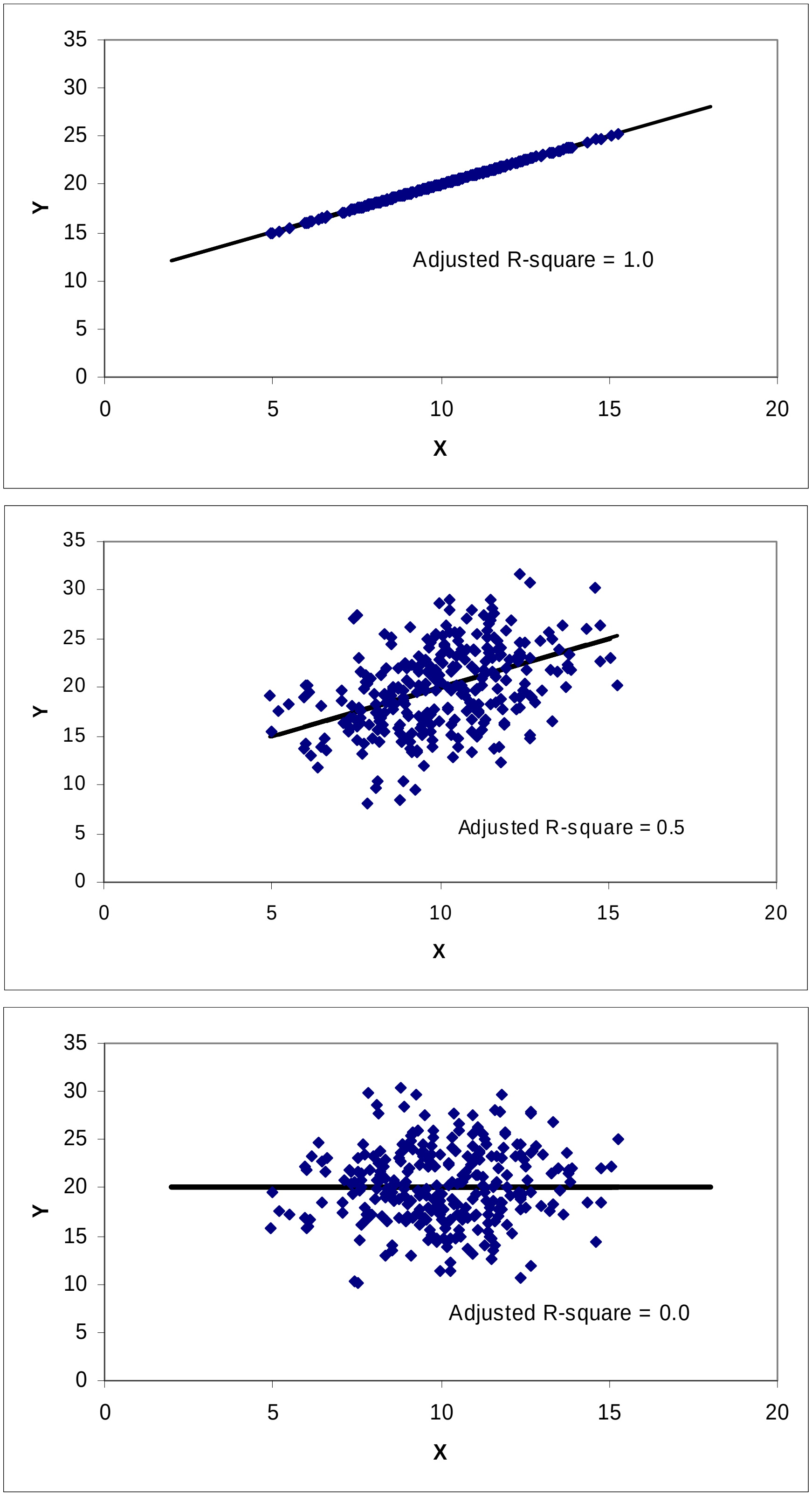
$$R^2\text{adjusted} = \frac{(\text{sample standard deviation of } Y)^2 - (\text{standard error})^2}{(\text{sample standard deviation of } Y)^2}$$

For the Lawrence Construction example, the adjusted R square is $(45{,}256^2 - 36{,}879^2)/45{,}256^2$, or 0.34.

In order to explain just what the adjusted R square tells us, we consider two extreme examples. First, consider the case where all the points fall exactly on the line (a perfect fit). In that case, the standard error will be zero and adjusted R square will be one. So an adjusted R square of one signifies a perfect fit. At the other extreme, suppose the model is of no help in predicting *Y*. In that case the line will be perfectly flat; $\hat{Y}$ will equal $\overline{Y}$ no matter what the value of *X* is; the standard error will approximately equal the standard deviation of *Y*; and the adjusted R square will be approximately zero. So an adjusted R square of zero signifies the model is of no help in predicting *Y*. See **Figure 4**.

Of course, most situations are somewhere in between these two extremes, and the adjusted R square will fall somewhere between zero and one. The higher the adjusted R square, the smaller the standard error *relative* to the standard deviation of Y. Whereas the standard error is an *absolute* measure, adjusted R square is a *relative* measure. Whereas the standard error carries the same units as the Y variable, the adjusted R square is scaled to be between zero and one.

Figure 4. Adjusted R-square scatter plots.



Finally, we offer one more way to interpret adjusted R square—as a percentage of variation explained by the model. You may remember that the square of any standard deviation carries the name variance. This means that the numerator in the adjusted R square calculation is the amount by which the variance is reduced if we use the model. The amount of variance reduction expressed as a percentage of the original variance (adjusted R square) is commonly referred to as the percentage of variation explained by the model.[9] The model in Lawrence Construction explains 34% of the variation is new-home construction cost. The variance in construction cost is $\$45{,}256^2$, 34% is explained by the model, and 66% ($\$36{,}879^2$) remains unexplained.

Two measures very similar to the adjusted R square are the *R square* and the *multiple R*. These latter two measures are clearly redundant in that the multiple R is the square root of R square. Because of this redundancy, we restrict our attention to R square.

R square is an "unadjusted" version of adjusted R square. Without getting into the gory details, adjusted R square "adjusts" R square downward to account for the number of independent variables used in the model. Adding variables to a model will always decrease the sum of squared residuals and thereby increase the (unadjusted) R square. But the adjustments built the standard error into the calculation, and the adjusted R square correctly account for the number of variables used. Adding variables to a model will not necessarily increase the adjusted R square. The difference between R square and adjusted R square is a function of the number of independent variables in the model relative to the number of data points. If there are lots of independent variables in the model and few data points, the adjusted R square will be much lower than the unadjusted. If there are a small number of independent variables and a huge number of data points, adjusted and unadjusted R square will be almost identical.

Standard error of the coefficients

We turn next to questions about the coefficients. Recall that we were careful to make a distinction between $\hat{b}$, the estimated coefficient, and *b*, the coefficient in equation (1). It should be clear that $\hat{b}$ is the slope of the fitted line (shown in **Figure 3**). It should also be clear that in the Lawrence Construction example, $\hat{b}$ equals 63.11. It might not be as clear what *b* is.

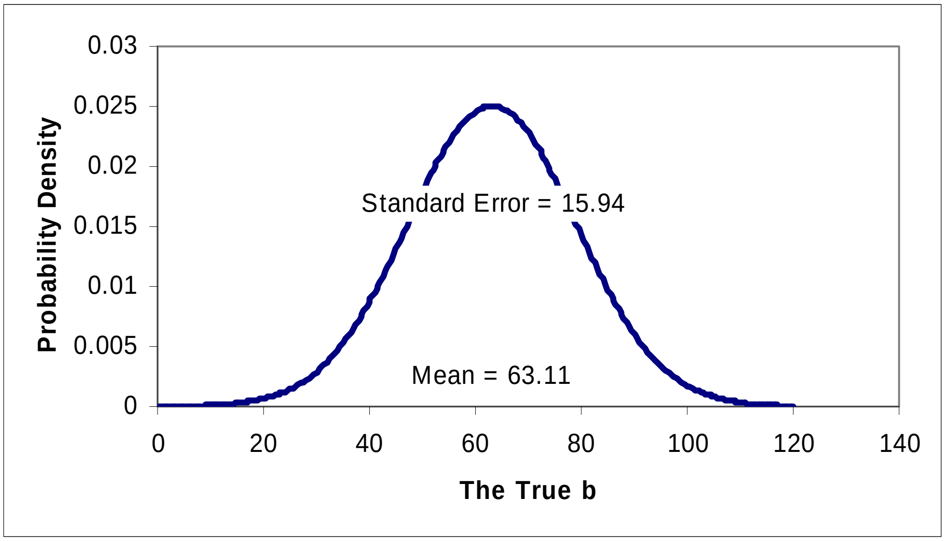
One way to think about *b* is as the slope of the imaginary line. The number *b* is the slope of the true underlying relationship between the mean of *Y* and *X*. Only with an infinite number of data points would we be able to know *b*.

Our fitted line is an estimate of the imaginary line (which has slope, *b*). Although $\hat{b}$ is our best estimate of *b*, it is not perfect. Luckily, we have some information about how close b-hat is likely to be to the true *b*. The measure of how close $\hat{b}$ is to *b* (and *b* is to $\hat{b}$) is the standard error

[9] Please be careful not to misinterpret adjusted R square as the percentage of time the model's predictions were correct. Because the model produces numerical forecasts of uncertain quantities, it makes no sense to talk in terms of correct versus incorrect predictions.

of the coefficient. In the Lawrence Construction example, the standard error of the $\hat{b}$ coefficient is 15.94. Our best estimate of *b* is 63.11 and the uncertainty surrounding that estimate has a standard deviation equal to 15.94. Our estimate of *b* is 63.11 plus or minus 15.94. **Figure 5** shows this forecast.[10]

Figure 5. Probability distribution for *b*.



The standard error of the coefficient allows us to make probability statements about *b* and answer almost any question we have concerning *b*. Most regression software packages anticipate questions about whether *b* is different from zero. We are often interested in whether *b* is different from zero because a *b* of zero means there is no relationship between *X* and *Y*. The t-statistic (which is simply the estimated coefficient divided by its standard error) measures the distance between zero and $\hat{b}$ in units of standard error. The higher the magnitude of the t-statistic, the further zero is from $\hat{b}$ and the more confident we are that *X* and *Y* are actually related. The p-value goes one step further and tells you the probability of seeing t-values greater in magnitude than the calculated t-statistic if *b* is actually equal to zero. In Lawrence Construction, if *b* is actually zero there is a 0.0005 probability of seeing a t-statistic with magnitude greater than 3.96. Because the t-statistic is so large and the p-value so small, we are very confident that *b* is greater than zero. Based on the data alone, we are very confident that construction cost is related to the size of the home. Put even more simply, the relationship between the cost of a new home and its size is statistically significant.

[10] Forecasts of model coefficient are normally distributed if the number of data points is large.

5. Assumptions behind the linear-regression model

Earlier in this note we showed how least squares can be used to fit the simple linear model to historical data. The resulting model can then be used to forecast the next occurrence of *Y*, the dependent variable, for a given value of *X*, the independent variable. This use of least squares to fit a forecasting model requires no assumptions. It can be applied to almost any situation, and a reasonable forecast results. At this level of analysis, least-squares modeling is equivalent simply to fitting a straight line through a cloud of points and interpolating or extrapolating for a new value of *Y* for a given *X* using the fitted line.

Although we need not make any assumptions to use this procedure, we leave an important question unanswered: How close can we expect the new *Y* to be to our forecast? Without some additional assumptions, we have no way of making a probability statement about the new *Y*. In many practical business situations, such a probability statement is an essential element in the decision-making process.

There is a procedure for measuring the uncertainty associated with a least-squares forecast that will produce a complete probability distribution for a new *Y*. This procedure brings real value and legitimacy to the regression modeling and forecasting process, changing it from a simple process—one step above graph paper and a ruler—to one that intelligently combines managerial judgment and statistical theory to produce believable point-and-interval forecasts.

That's the good news. The inevitable bad news is that in order to make probability statements about a new *Y* using a least-squares regression model, a variety of assumptions must be made. In other words, probability statements made using linear-regression theory are true only if certain assumptions hold. You can thus see the importance of (1) understanding these assumptions, (2) knowing how to check their validity, (3) understanding the consequences of an incorrect assumption, and (4) knowing what can be done if the assumptions do not hold. This note addresses each of these four points for the four general assumptions behind linear regression. The model must be checked for (1) linearity, (2) independence, (3) homoskedasticity, and (4) normality.

For simplicity and convenience, most of this note will refer to the simple linear-regression model. The material presented, however, can be applied as well to the general linear multivariate regression model.[11]

Linearity

The first and most fundamental assumption behind simple linear regression is that the model does indeed fit the situation at hand. The expectations for the *Y* variable, therefore, are linearly related to the value of the *X* variable. The equation expressing this relationship was given earlier as equation (2)

[11] The multivariate linear-regression model expresses *Y* as a linear function of more than one independent variable.

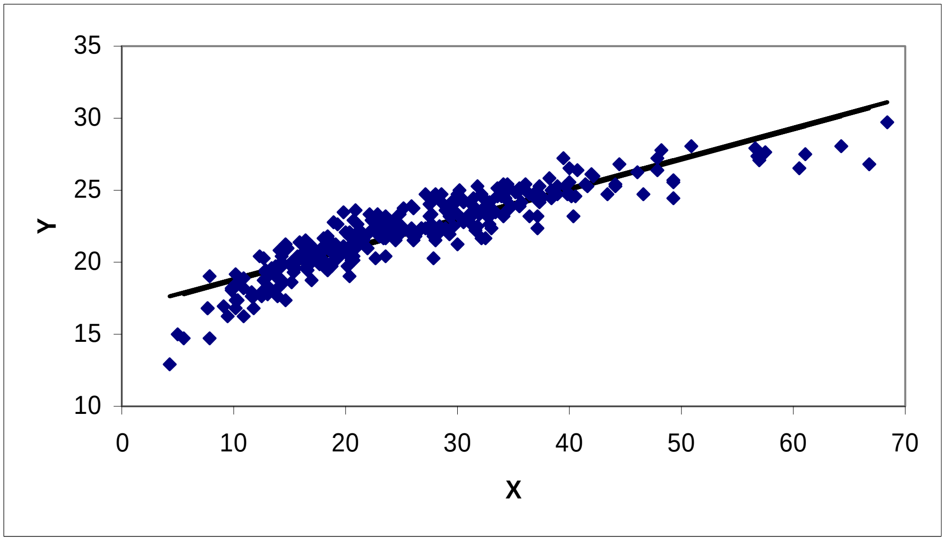
$$\text{Mean of } Y \big| X = a + bX,$$

which tells us that the mean value of *Y* changes in a linear manner with *X*.

As a further explanation of this assumption, recall that the mean of an uncertain quantity is the long-run average value expected from an infinite number of trials. Thus, if we held *X* constant, observed *Y* a great many times, and then calculated the average of the *Y*s, we would get a value pretty close to the mean of *Y*. Equation (2) states that if we did the same thing for a variety of different *X* values, a plot of the means of *Y* versus *X* would be a straight line with slope *b* and intercept *a*.

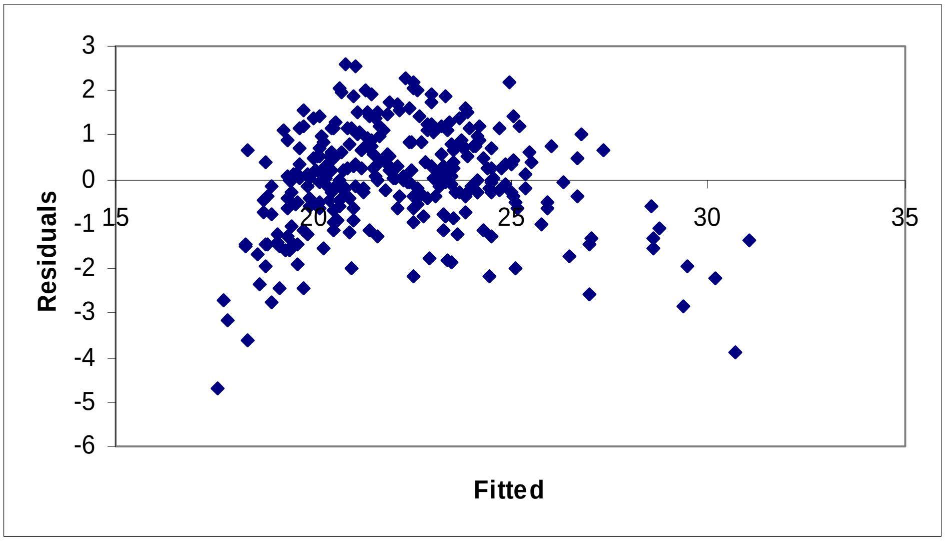
This suggests a very simple way to check to determine if the simple linear model fits the situation at hand. A scatter plot of *Y* versus *X* values should show a pattern that looks linear. **Figure 6** shows a situation for which this assumption definitely does not hold. Notice that the points tend to be below the line at either end and above the line in the middle. In this case, we say that the linear model exhibits *lack of fit*.

Figure 6. Scatter plot showing nonlinearity.



A second way to check for linearity is to examine a scatter plot of residuals versus fitted values for the least-squares regression line. This method is particularly useful when the fitted model contains several *X*-variables. If the model fits the data, the residuals should appear to have a mean of zero for all values of *X*. **Figure 7** is a scatter plot of residuals versus fitted for the data in **Figure 6**. Notice that these residuals tend to be negative at the ends of the plot and positive in the middle—indicating an inadequacy in the form of the fitted model. Inadequacies in form show up as frowns (**Figure 7**) or smiles in a scatter plot of residuals versus fitted values.

Figure 7. Scatter plot of residuals versus fitted showing nonlinearity.



In addition to these simple checks for lack of fit, there are statistical tests designed to measure the degree to which the proposed model fits the observed data.[12]

Using the least-squares model for forecasting requires considerable judgment on the part of the model builder who must be willing to assume that equation (2) is reasonable over the relevant range of interest. In any case, the modeler should proceed only with the understanding that the accuracy of his or her forecasts is predicated on the assumption that the model is indeed correct.

What are the consequences of employing a model that does not fit the situation? Quite simply, the model gives inaccurate forecasts. Looking again at **Figure 6** or **Figure 7**, we see that the forecasts would be too high when *X* is either very low or very high and would be too low if *X* is close to the average.

What can be done if a prospective model exhibits a significant degree of lack of fit? Find a better model, probably one with a different functional form. After studying **Figure 6** or **7**, for example, the model builder might assume that the mean value of *Y* is linearly related to the square root of *X* rather than to *X* itself. The resulting model,

$$\text{Mean of } Y\,|\,X = a + bX^{½},$$

would still be a simple linear-regression model (the new independent variable is now the square root of *X*) that could fit the situation better than the old model.

[12] For an explanation of one such test, see Dielman, *Applied Regression Analysis for Business Economics* (Boston: PWS-Kent, 1991), 196–197.

The remaining three assumptions all deal directly with the errors in the model. Recall that the complete specification of the simple linear model given earlier as equation (1),

$$Y = a + bX + \text{error},$$

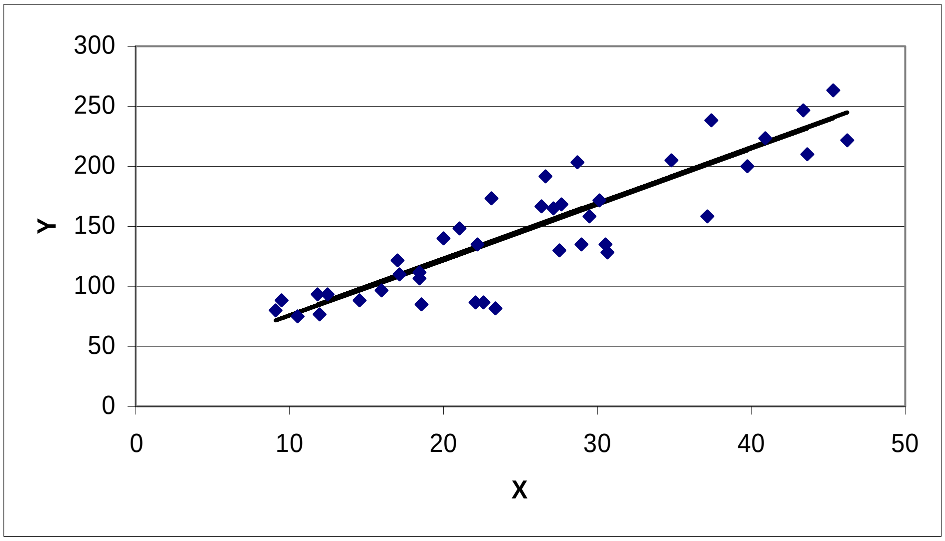
includes an error term representing the uncertainty inherent in the situation. We assume that the mean of *Y* varies linearly with *X* and, furthermore, that each *Y* will vary consistently and randomly about its mean. The remaining three assumptions specify what we mean by *consistently* and *randomly*.

Independence

The second assumption is especially important in those situations where there is a definite time ordering of the *X* and *Y* observations. It states that error_i and error_j, the uncertain portions of *Y* for any two trials, must be uncorrelated—that is, the outcomes from different trials must be independent.

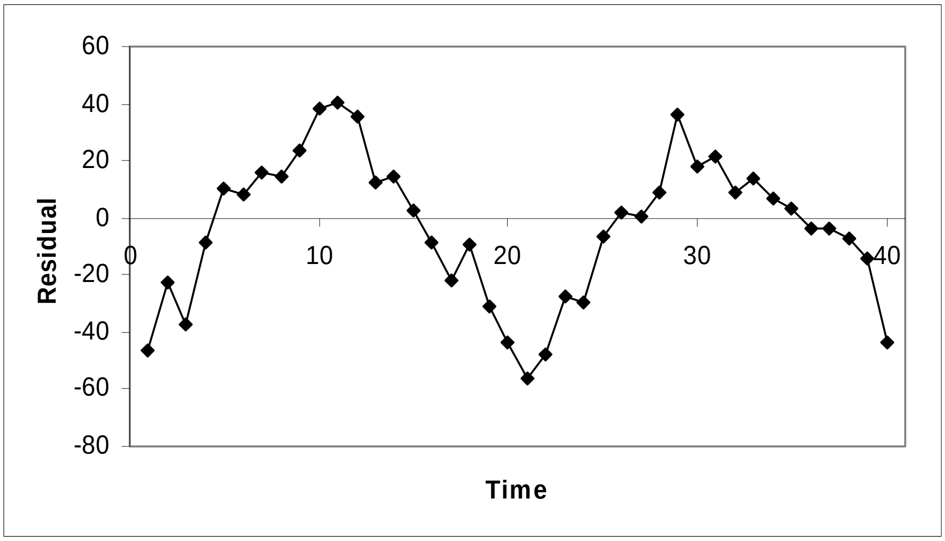
Figure 8 and **9** show a situation in which this assumption is violated. From the scatter plot of *Y* versus *X* in **Figure 8**, we see that the simple linear model fits the data quite well. Judging from this figure, everything looks fine. If, however, we plot the errors from the model in their correct time order (**Figure 9**), we see a definite pattern: a run of four negative residuals followed by eleven positive, ten negative, ten positive, and five negative. These long runs indicate that the errors are not random across time. This pattern indicates that these errors do not satisfy the second assumption: independence. Because these errors are correlated with each other, we say they are *autocorrelated*.

Figure 8. Scatter plot of *Y* versus *X* not in time order.



What are the consequences of using the model when there is temporal dependence or autocorrelation within the errors? Again, the model will give poor forecasts. Looking carefully at **Figure 9**, we see that the time pattern of the residuals strongly suggests that the residual for the next trial, trial 41, will be negative. If we used the model to forecast observation 41, the forecast would likely be too low. Ignoring the autocorrelation, therefore, would lead to forecasts that are not as accurate as they could be. A secondary consequence of autocorrelation is that the least-squares estimates of the model coefficients are not as accurate as they could be.

Figure 9. Time series plot of residuals.



Perhaps the easiest way to check temporal independence is to plot the least-squares residuals from the fitted model in time sequence (as in **Figure 9**). A visual inspection can often reveal patterns suggesting that the independence assumption is violated. If the plot appears to represent a completely random sequence, however, the model builder should accept the assumption of temporal independence. Because this visual inspection can sometimes be misleading, however, a variety of tests are available that specifically address the hypothesis of independence across time.[13]

What can be done when significant autocorrelation presents itself? Again, we can only suggest that the model be improved to explain the cause of the autocorrelation. Often the model builder must find a new independent variable or use lagged versions of current variables.

Homoskedasticity

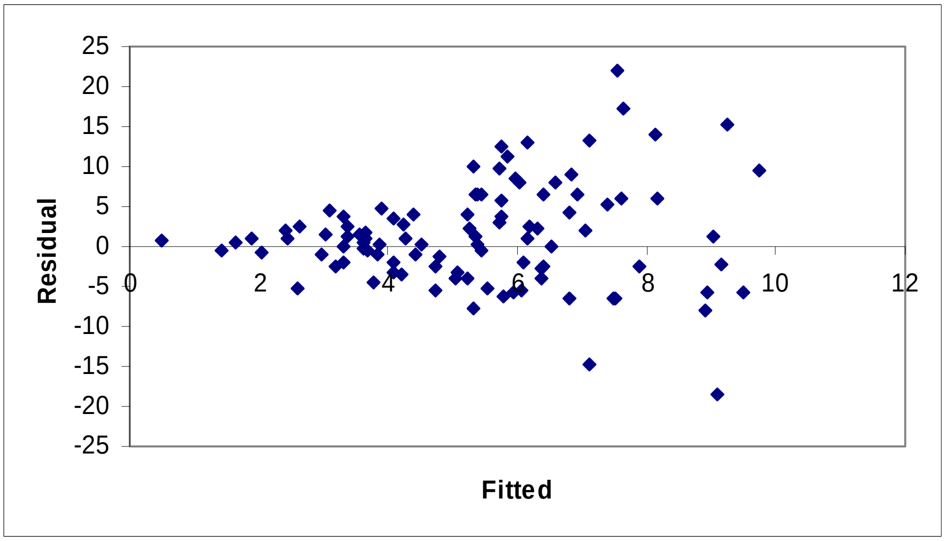
Homoskedasticity is a big word with a simple definition: constant variance. The homoskedasticity assumption in a model means that the variance or scatter of the errors does not

[13] For a description of statistical tests for temporal independence, see Dielman, *Applied Regression Analysis for Business Economics*, 271–277.

change from trial to trial. By implication, there is a constant amount of uncertainty surrounding each *Y*, regardless of the value of *X*.

Figure 10 depicts a situation where *heteroskedasticity*—nonconstant variance—is present. Note that the variance or scatter of the *Y*s about the line increases with *X*. This situation is common. Uncertainty often increases with the size of the dependent variable. A second (less likely) way for heteroskedasticity to occur is across time—that is, when the variance of the errors increases or decreases with time (as shown by a time-series plot of the residuals).

Figure 10. Scatter plot of residuals versus fitted showing heteroskedasticity.



How does heteroskedasticity affect our forecasts? Although the point forecasts themselves are not necessarily affected, heteroskedasticity does lead to a misspecification of the uncertainty surrounding a forecast. For the example in **Figure 10**, linear-regression theory would use an average measure of uncertainty (the standard error) over the entire relevant range. Consequently, we would underestimate the amount of actual uncertainty present in high-valued forecasts and overestimate the amount present in low-valued forecasts. Thus, heteroskedasticity can be critical in making decisions that require both a point and an interval forecast.

The presence of heteroskedasticity also indicates that least-squares coefficient estimates are not quite as good as they could be. Better estimates are obtained if more weight is put on those observations that are more certain (have lower variance) relative to those observations that are less certain. Because least squares weights each observation equally (it minimizes the total sum of squared residuals), the high-variance observations exert too much influence on coefficient estimates.

We offer a pair of alternatives for dealing with heteroskedasticity. One approach is to transform the *Y*-variable, changing both its units and its interpretation. Instead of using the original *Y*-variable as the dependent variable, a model builder might try using *Y/X*, the square root of *Y*, or the reciprocal of *Y*. If heteroskedasticity had been a problem in the Lawrence Construction model, for example, one might try modeling cost-per-square-foot (*Y*/*X*) rather than dollar cost. The second alternative requires using a more sophisticated estimation procedure called weighted least squares.[14] Essentially, the model builder replaces the assumption of homoskedasticity with one that explains exactly how the variance of the errors changes with *X*.

Normality

The final assumption implied by using the model is that the errors are normally distributed. Because these errors are the net effect of a very large number of unmeasurable or unexplainable factors that exert influence on the dependent variable, it is often reasonable to assume that the sum of a large number of small uncertain quantities follows the normal distribution.

Consequently, if the error is normally distributed, then *Y* for a given *X* will also be normal. This assumption is important because once we know the probability distribution of the dependent variable, we can easily make probability statements. All we need in order to use the normal distribution is the mean and the standard deviation.

One easy way to check this assumption is to plot a histogram of the residuals from the model. If the plot approximates the well-known bell-shaped normal curve, the assumption of normality is probably reasonable. **Figure 11** shows a histogram of residuals that suggests nonnormality. More sophisticated ways to test normality include the normal probability plot[15] and the chi-squared goodness-of-fit test.[16]

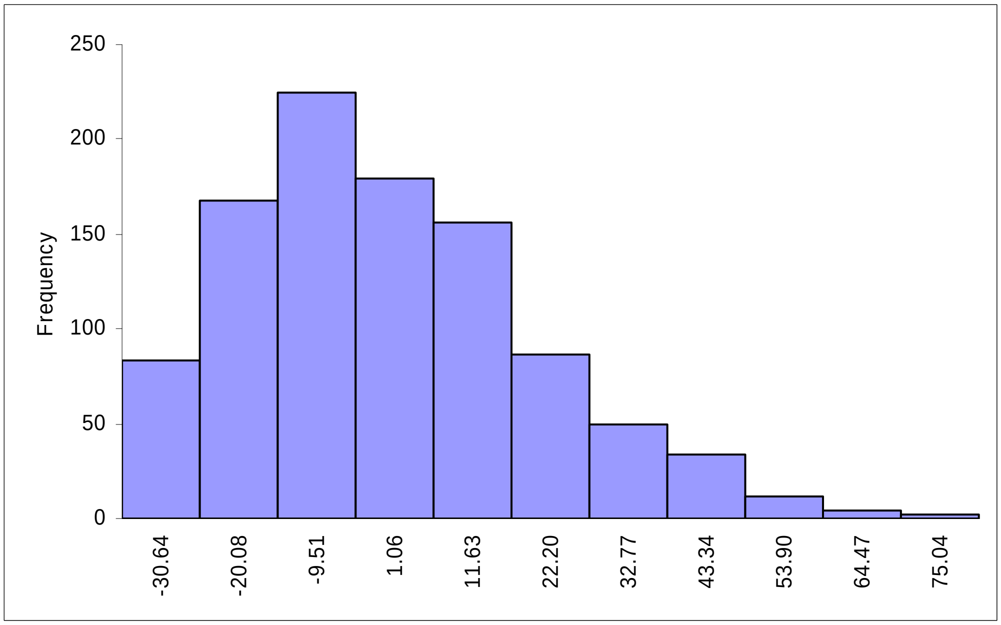
The consequences of an incorrect assumption of normality are not that severe. For example, if we used the normal distribution to make a probability statement for the example in **Figure 11**, our biggest mistake would be in underestimating the probability of a very large error. This inaccuracy would occur because the distribution of errors is skewed, but the normal is symmetric.

[14] Details of weighted least squares can be found in most texts on regression analysis or econometrics (e.g., Draper and Smith, *Applied Regression Analysis*, 2nd ed. [New York: Wiley, 1981], 108–116).

[15] Sort the residuals, assign sample cumulative probability k/(n + 1) to the kth ranked residual, calculate the standard normal values for each of these probabilities (use =NORMSINV in Excel), scatter plot these standard normal values versus the sorted residuals. If the residuals are normal, the plot should be a straight line.

[16] Details of the chi-squared goodness-of-fit test can be found in most statistics texts. The test compares the observed cell frequencies of a histogram such as in Figure 11 with the frequencies expected if indeed the observations came from a normal distribution.

Figure 11. Histogram of residuals showing nonnormality.



One possible solution to a problem of severe nonnormality is to take the logarithm of the original dependent variable. Doing so will change the nature of the Y-variable, but may produce a new dependent variable that satisfies the normality assumption. Another solution is to abandon the normality assumption in favor of a more appropriate probability distribution for the errors. Unfortunately, this latter solution is difficult to implement.

Summary of regression assumptions

We have seen that in order to make intelligent use of the simple linear-regression model in decision making, we must realize what assumptions we are making. Most of these assumptions can be stated using the error term in the model

$$Y = a + bX + \text{error}.$$

These assumptions are

- Linearity (the mean of *Y* must be a linear function of *X*)
- Independence (the errors must be uncorrelated with each other)
- Homoskedasticity (the errors must have constant variance)
- Normality (the errors must be normally distributed).

To check these assumptions we usually begin by examining the residuals from the least-squares regression model. **Table 4** summarizes how to use residuals and residual plots to check each assumption. If the assumptions hold, the residuals should appear completely random and without pattern, no matter how they are viewed. Two very important ways to look at these residuals are to plot them in time sequence (checking for autocorrelation and, perhaps, heteroskedasticity) and scatter plot them against the fitted values (checking for lack of fit and heteroskedasticity). If inadequacies are found, this note has suggested alternatives for correcting—or at least improving—the situation. **Table 4** summarizes these proposed remedies. We have also outlined the consequences of employing the model when these assumptions are violated.

Table 4. Diagnostic Checking of Regression Assumptions.

Assumption	**Diagnostic Checking**	**Possible Remedy**
Linearity	Examine scatter plot of residuals versus fitted ($\hat{Y}$) for evidence of nonlinearity.	Try transforming *X*-variables or including new *X*-variables which explain/eliminate the nonlinearity.
Independence	Plot residuals in time order and look for patterns.	Include lagged variables in the model.
Homoskedasticity	Examine scatter plots of residuals versus fitted ($\hat{Y}$) and residuals versus time and look for changing scatter.	Transform the *Y*-variable.
Normality	Examine histogram of residuals. Look for departures from normal bell-curve shape.	Transform the *Y*-variable. Use a distribution other than the normal when making probability statements.

6. Model-building philosophy

Linear (regression) models are used in a variety of business situations for a variety of purposes. One reason for their popularity is that they can be easily understood and implemented. All that is needed is data on two or more variables and a computer or calculator. With a minimal amount of effort, a manager obtains: (1) an equation for the relationship between the dependent variable (*Y*) and one or more independent variables (the *X*s), (2) a forecast of *Y* for any given set of *X*-values, and (3) a host of statistics to complement the analysis. Because so much is obtained so easily, the potential for misuse is great if the model builder is not careful about how the model is built and applied.

This section provides guidelines for the effective construction and application of linear models. We begin with a discussion and categorization of the various *uses of the linear model* (description, forecasting, and control). Whether a given model is appropriate for a particular use depends upon the *nature of the relationship among the variables* (chance, correlation, or causal). We give particular attention to illustrating *the importance of the underlying relationship to the use of the model*.

Even when a manager understands what kind of relationship is necessary for the particular way the model will be used, there still remains the question of how best to combine this manager's knowledge of the situation with available data to build an effective model. We propose a *model-building procedure* for building a linear model when there is only limited prior knowledge about the underlying relationships among the variables. The procedure attempts to maximize the use of the managers' prior knowledge and judgments, yet allows for the constructive and intelligent use of the available data. The objective is to guard against overreacting to peculiarities of the data and to build a sound and effective model that is consistent with both the modeler's prior knowledge and the information found in the data. In contrast to this positive advice on how to go about building a good model, we end this section with a list of things not to do, explaining some *common mistakes* associated with the building and using of linear models.

Uses of the Linear Model

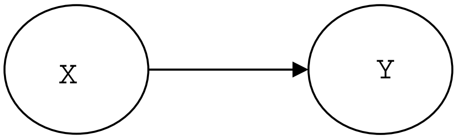
Description is the least ambitious of the three uses. The purpose of the model is merely to describe and measure the historical relationship among two or more variables. Although the resulting model often has no direct or immediate impact on a decision, the knowledge gained implicitly affects future decisions. For example, if a regression model were used to describe the relationship between salaries and work experience of a group of employees, the only reason for constructing such a model would be to better understand the relative importance the company has placed on longevity or seniority in its salary practices.

A second use of the linear model is for *forecasting* the variable of interest (the dependent or *Y*-variable) for a given set of known, independent *X*-variables. The resulting forecast can be a valuable aid to decision-making. For example, an office-cleaning company employs a linear model, relating the time to clean an office building to the physical characteristics of the building (size, number of windows, number of offices, etc.). A forecast of how long it will take to clean a new office building is then used to decide how much to charge for cleaning services.

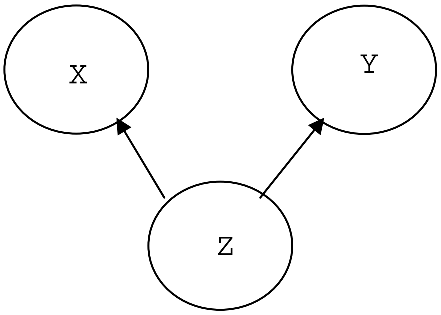
The most ambitious use of a linear model is *control*. Once the relationship between *X* and *Y* is established through a linear model, the decision-maker then might use the model to decide how best to change the value of *X* to effect a desired change in *Y*. An example is a marketing-mix model relating the sales of a product to its price, advertising expenditures, and levels of distribution. The model is used for control if its purpose is to determine the optimum level of advertising. Here the manager changes *X* (advertising expenditures) to induce a change in *Y* (sales). Control is more ambitious than forecasting, because of the requirements it places on the nature of the underlying relationship between the dependent and independent variables.

Nature of the relationship among variables

The strongest relationship between two variables is *causal*. If the value taken by the dependent *Y*-variable actually depends directly on the value of the independent X-variable, we say that *X* causes *Y*. For example, *X*, the level of advertising, causes *Y*, the level of sales. If we changed *X*, we would expect to see a change in *Y*.



Variables *X* and *Y* are positively *correlated* if high values of *X* tend to occur with high values of *Y* and vice versa. Negative correlation between two variables implies that high values of *X* occur with low values of *Y*. Obviously, if *X* causes *Y*, then *X* and *Y* are correlated,[17] but the opposite is not always true. Variables *X* and *Y* could be correlated because they both are caused by a third lurking variable, *Z*.



For example, *X* and *Y*, a student's scores on two separate tests of intelligence, would be correlated because both are caused by a lurking third variable, actual intelligence. Clearly, however, the student's score on one test does not cause the score on the second test.

We use this last category to refer to those situations in which no actual underlying relationship exists between *X* and *Y* even though the available data strongly suggest otherwise. *Chance* relationships are sure to be found if the model builder blindly dredges through a large number of independent variables in an attempt to find one that shows a strong historical relationship with the dependent variable of interest. If enough different *X* variables are considered, one or more will always be found that appear to be significantly related to *Y*.

A well-known example is that a relationship has been found to exist between the league that produced the World Series winner and the political party of the person elected president of the United States. Perhaps one can never be absolutely certain that there is not some underlying relationship between baseball and politics, but the much more plausible explanation for this observed relationship is that many people looked at many possible ways to predict winning candidates and this relationship was found and reported. Most likely it occurred by chance alone.

[17] There are, however, pathological examples of variables that are causal, but not linearly correlated.

The importance of the underlying relationship to the use of the model

If the eventual purpose of the model-building effort is to control, to change *X* in order to induce a desired change in *Y*, then it should be clear that the underlying relationship must be causal. If advertising causes sales, a change in advertising expenditures will change sales. In contrast, if the underlying relationship is only correlative, changing *X* does not change *Y*. In the example of the two IQ tests, if we increase a student's score on the first test (by giving the student the questions beforehand or simply adding l0 bonus points to the score), we do not expect to produce a change in the score the student will get on the second test.

The only way to prove and measure causality is by experimentation. To see if a change in *X* will change *Y*, we actually have to change *X* and observe what happens to *Y*. If we rely only on available data (sometimes referred to as happenstance or unplanned data), it is very difficult to distinguish causation from mere correlation.

For example, the observed correlation coefficient of +0.90 between the number of stork nests counted in the city of Stockholm, Sweden, and the number of babies born over several years in the l930s and 1940s, did not prove, as some might suggest, that storks were responsible for babies. One explanation is that the racket made by the storks in the early-morning hours as they left their nests caused the inhabitants of the houses below to wake up, have trouble getting back to sleep, and … (you can figure out the rest). An alternate and more plausible explanation is that the '30s and '40s were a period of affluence in Stockholm—with considerable home construction providing new nesting places for storks as well as living space for the expanding population. The point is that a strong correlation between *X* and *Y* does not necessarily imply that *X* causes *Y*. Thus, in order to use a linear model for the purposes of control, the modeler must either experiment with *X* and observe the change in *Y* or rely on some prior knowledge or theoretical considerations as a basis to assume that *X* causes *Y*. Otherwise, we might find ourselves shooting storks to slow down the population growth of Stockholm.[18]

If the purpose of the model-building effort is forecasting, causation is not necessarily required and correlation can suffice. The student's score on the first IQ test can certainly be used to forecast the score on the second. Similarly, it might even be possible to use the number of stork nests to forecast the number of babies. One note on using *X* to forecast *Y*: *X* must be known in order to use it in the model to provide a forecast of *Y*. Thus, to be of direct use in forecasting, the stork/baby model would have to relate the number of babies born during a period to the number of stork nests at the *beginning* of the period.

Finally, if the purpose of the linear model is only to describe the observed historical relationship among two or more variables, its use is always appropriate. Even if there is no underlying relationship between *X* and *Y*, it is correct to use the linear model to make a statement about what was observed in the data.

[18] This and other examples of faulty inferences can be found in Robert E. Fuerst, "Inference Peddling," *Psychology Today* (March 1979): 92–95.

The following table summarizes the nature of the underlying relationship necessary for each of the three uses of the linear model:

Model Purpose	**Relationship Required**
Control	Causal
Forecasting	Causal or Correlative
Descriptive	Causal or Correlative or Chance

We can look at the three uses of the linear model as three kinds of statements we wish to make about *X* and *Y*.

The weakest statement is simply a description of an observed relationship. For example, if the variables of interest are yearly sales and advertising expenditures, then the following is an example of a purely descriptive statement about their relationship:

The best (least-squares) linear relationship
between sales and advertising expenditures was
Sales = $400,000 + 2 × Advertising
based on the previous 20 years of annual data.

Note that this statement requires no assumption about the underlying relationship; it is merely a statement about what happened.

A slightly stronger statement is

Since advertising expenditures for next year will be
$l00,000, we expect sales to be $600,000.

This forecasting statement is appropriate only if we assume that the historical relationship is representative of some underlying relationship that will continue unchanged next year. We must assume that advertising and sales are correlated.

The strongest statement to be made is

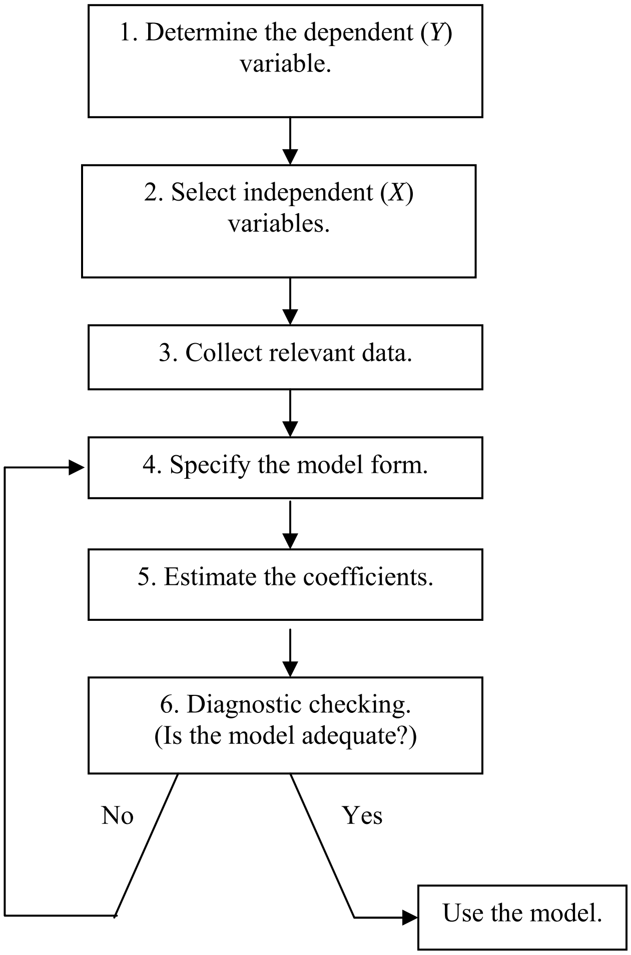
If we increase our advertising by $l,
we expect an increase of $2 in sales.

In order to make this statement, we must assume that advertising causes sales, and that the causal relationship continues unchanged next year.

Model-building procedure

Figure 12 outlines a six-step iterative procedure for building a linear model. The purpose of this procedure is to combine intelligently and carefully the modeler's knowledge of the situation with the information available in the data to produce a useful model.

Figure 12. Interactive model-building procedure.


1. Determine the dependent (*Y*) variable.
2. Select independent (*X*) variables.
3. Collect relevant data.
4. Specify the model form.
5. Estimate the coefficients.
6. Diagnostic checking. (Is the model adequate?)
No
Yes
Use the model.

The important thing to keep in mind is that modeling is not an exact science. When drawing conclusions from the data, the model builder must be careful to use sound judgment and to rely heavily on prior knowledge of the situation. It is quite acceptable to study the available data and learn from it, but the model builder must be stingy with its use to avoid being misled into believing in a relationship that happened in the past, but that will probably not happen in the future.[19]

l. Determine the dependent variable (*Y*)

The purpose of the modeling effort usually dictates what quantity is of interest. Exactly how to define the dependent variable, however, is not. Suppose we wish to forecast sales. Should it be sales in units or in dollars? Should we use constant dollars or current dollars? Should it be total sales or our share of the total market? Should it be our sales or industry sales? Should it be sales for a month, quarter, or year? Maybe we should look at change in sales. By sales, do we mean shipments, orders, or demand?

In making the choice of dependent variables, three criteria should be considered. How *useful* is the dependent variable to the purposes for which the model is being built? A forecast of sales for next quarter may be more useful than a forecast of sales for next year if we update our production schedule quarterly. Second, how *attainable* is the dependent variable that is needed? Some dependent variables are easier to collect than others. Number of shipments is probably much easier to measure and obtain than total number of units demanded, which is almost impossible to measure because it is meant to include lost sales. The last criterion is how *modelable* the variable is. This consideration ties in with the considerations of step 4, determining the functional form of the model. The dependent variable is but one side of the model equation, and as such it must be chosen to make sense when coupled with the independent variables. For example, whether to use sales or change in sales depends on which makes better sense in light of the available independent variables.

2. Choose the potential independent variables

The primary concern in choosing the list of potential independent (*X*) variables is that they make sense. The model builder should consider only those variables for which there is reason to believe (from prior experience or theory) that they have an effect on the dependent variable. It should be kept in mind, however, that even as a blind hog sometimes finds an acorn, a long list of nonsensical *X*-variables will be sure to include one or more that show a strong historical relationship to *Y*. In order for such a finding to have meaning (i.e., to be due to an actual relationship rather than chance), the list of *X*-variables must be limited to those that have been preselected as reasonable.

In addition to making sense, the attainability and usefulness of a variable should also be considered. If the model is to be used for forecasting, the independent variables must be known (or at least more predictable than the dependent variable) in order to be functional. Similarly, the

[19] This admonition to be stingy with the use of data assumes we are working with a small data set. With a large data set, the model builder has the luxury of setting aside or holding out a subset of the data for later use in validating a candidate model. Using this split-sample technique, the model builder can be more aggressive in exploring the data in the modeling set knowing that the candidate model will be tested on the holdout sample. For additional details on this split-sample technique, see Kleinbaum, et al., *Applied Regression Analysis and Other Multivariate Models*: 2nd ed. (PWS-Kent, Boston, 1988), 328–331.

decision maker must be able to manipulate the independent variable if it is to be used to control the dependent variable.

3. Collect relevant data

The key to this step is to collect data that are from similar situations. Each observation must be similar to all others in the sense that the underlying relationship between the *X*s and *Y* must be the same in each case. It is expected that the *X*s change from observation to observation, but the relationship (i.e., the coefficients of the underlying model) must remain constant. Thus, the data must come from situations that are indistinguishable from each other in the sense that the expectation of *Y* is a function only of the *X*s, and not of some other factor. If another factor exists that can be used to distinguish the data, that factor is a candidate for an independent variable. A second way to handle distinguishable data is to separate the data into smaller sets that can reasonably be considered to represent similarly indistinguishable situations.

4. Hypothesize the model form

This step is probably the most important and requires the most judgment. The model builder must attempt to incorporate his or her knowledge of the situation (based on experiences and available theory) into a single algebraic equation. The equation should be as simple as possible, yet remain consistent with the modeler's knowledge of the important complexities of the situation.

Simpler models are generally better models for two reasons. First, the simpler the model, the less it depends on the peculiarities and uncertainties of the limited data set to which it was fit. The more complicated the model, the better it will fit the data, but the more likely it will be to go wrong sometime in the future. A second reason for preferring a simple model is that it will be easier to use.

The desire for a simple model, however, conflicts with the importance of building a model that is entirely consistent with the manager's knowledge of the situation. Each model form implicitly makes certain assumptions about the way things work. The modeler must scrutinize a candidate model to be sure he understands its implications and assumptions and agrees with them.

As an example, if *Q* is the quantity demanded of a given product and *P* its price, three reasonable functional forms for the relationship between *Q* and *P* are:

$$Q = a + b\,P \tag{4}$$

$$Q = a + b(1/P) \tag{5}$$

$$ln(Q) = a + b\,ln(P) \tag{6}$$

All three of these are relatively simple, though an argument can be made that (4) is the simplest.

A careful look at these models clarifies the differences in the assumptions behind each. Model (4) says that an increase in *P* of 1 unit will change *Q* by *b* units at every level of *P*. Therefore, we probably expect *b* to be negative, but wonder if the effect of price on quantity can be the same no matter what the absolute level of price. In addition, the model says that when $P = 0$, *Q* will equal *a*. Also, for a *P* large enough, *Q* will be negative. Obviously, this model can only be good over a limited range of prices. Model (5) says that a unit increase in price will change *Q* by an amount that decreases with the level of price, perhaps something that makes a little more sense. In addition, because *b* will be positive if *Q* decreases as *P* increases, model (2) says that as *P* decreases to zero, *Q* becomes extremely large. Likewise, as *P* becomes very large, *Q* approaches the value *a*. Thus, we would expect *a* to be about zero. Model (5) may make more sense than model (4), and it is definitely better if the range of prices used is large.

Finally, model (6) may appear to be weird, but it does have a very important property. It says that a percentage change in price produces the same percentage change in *Q* at all levels of price. This property is called constant elasticity. The elasticity, or ratio of the percentage change in *Q* to the percentage change in *P*, is measured by *b* in this equation.

In summary, each model makes a different statement about the relationship between *P* and *Q*. We stress the importance of recognizing the assumptions and implications behind a candidate model form and judging their appropriateness prior to fitting the model to the data.

5. Fit the model

This is perhaps the easiest of the steps. Ordinary least squares is used to select the values of the model coefficients that minimize the sum of squared errors between past actual *Y* values and model predictions. Details of least-squares estimation were given in Section 2.

6. Perform diagnostic checking

There are two steps in checking the proposed model. The first is to make sure that all of the assumptions required for the use of a linear model (homoskedasticity, lack of autocorrelation, etc.) are met. Section 5 addressed these issues and suggested ways to change the model (in effect, a return to step 4) if inadequacies are found.

The second aspect of diagnostic checking is testing the significance of the model coefficients. Consistent with the desire to build the simplest possible model and to avoid overreacting to the peculiarities of the data is the importance of checking to make sure that the estimated coefficients are significantly different from zero. If a coefficient is not very different from zero (as measured by the t-statistic and p-value associated with that coefficient), the model builder should seriously consider removing the associated *X*-variable from the model. The new model will be simpler and will fit the data almost as well.

Because we know that adding any independent variable to the model will improve the fit and show some measure of significance, it is important to be conservative when choosing those variables to remain. The final model should include only those variables that we are convinced are actually related to the dependent variables. This conviction comes from both our knowledge of the situation and the observed relationship in the data. Dropping an insignificant variable from the model is a return to step 4.

Some caution must be exercised as variables are added and deleted from a model. Always keep in mind that if a single variable is added to or deleted from a model, the relative importance of all other variables changes. Thus, if two coefficients in a candidate model are not significant, it is a good idea to remove them one at a time. It can happen that once the first variable is dropped, the other may then appear significant. This often happens if the two *X* variables in question are correlated with each other. The low significance of both t-statistics means that both *X*s do not need to be in the model together. It does not necessarily mean that neither one belongs in the model.

Common mistakes

In contrast to all that has been said about how to build a model, we close with a list of things *not* to do. By no means an exhaustive list, the mistakes that follow are the most common errors made in building and using linear models.

The fishing expedition. The modeling effort begins with a brainstormed list of all the independent variables that could conceivably be related to the dependent variable. Data are gathered on as many of these as possible, and the modeler spends the rest of the time weeding through the mass of numbers. Just as it is no surprise that a diligent fisherman eventually catches something, it is also quite clear that such a procedure will probably produce a model that fits the data quite well. The important question, however, is whether the resulting model fits well because it is indeed an accurate representation of some underlying relationship, or because it happened by chance alone to be the best-fitting model of the very many tried.

More is better. If a simple one- or two-variable model works well, will not a four- or five-variable model work even better? Although it is true that making a model more complicated will always improve the fit, it is quite another matter to assume that the better the fit, the more useful the model. An incredibly complicated model will certainly conform to the peculiarities of the available data and fit the past quite well, but it will also probably go wrong when used to predict the future.

Forecasting with a forecast. No matter how strong the relationship between *X* and *Y*, the ability to use *X* to forecast *Y* is predicated on knowing *X* at the time the forecast is needed. It does no good to build a model to help forecast automobile sales as a function of gasoline prices if gasoline prices are just as hard to forecast as automobile sales.

Correlation as causation. A strong linear relationship between *X* and *Y* does not necessarily mean that if you change *X*, you can cause a corresponding change in *Y*. Don’t kill storks as a means of birth control.

What was the R^2? The percentage of variance explained (R^2) is often used as the single measure of the quality of a linear model. There are two good reasons why you should not always be impressed with a high R^2: (l) complicated models and small data sets make it easy to get a high R^2, and (2) if both *Y* and *X* exhibit a trend across time (as do almost all business and economic variables), then a high R^2 is to be expected, irrespective of any real relationship between the two variables. In addition, R^2 only attempts to check the quality of fit of the regression, and we all know that a host of other things must be considered (residual plots, t-statistics, etc.) when judging a model.

Extrapolation—a leap of faith. Users often forget the assumptions that must be made in order to apply the model when forecasting a new situation. Foremost among them is the assumption that the underlying relationship specified by the model does indeed hold for the new situation. The new situation can be different only in that it involves a new set of *X*-values (that's the whole point of regression), but everything else (i.e., the relationship between the *Y* and *X*s) must remain the same in order for the model to work effectively.

Summary

This section has briefly described the various uses of a linear model (description, forecasting, and control) and the types of relationships between variables (causal, correlative, and chance). A table summarizes the kinds of relationships necessary to each of these three uses.

The remainder of this section discussed how to build a good model. The main idea was that model building is much more than fitting curves to data. The procedure put forth called for a majority of effort to be spent in analyzing the situation, selecting variables, and specifying a form for the model—all prior to using a computer to calculate coefficients and fit curves. The iterative procedure then called for a judicious use of the data to estimate coefficients, check assumptions, test the significance of proposed variables, and refine ideas about how things work.

7. Forecasting using the linear-regression model

In the previous section, we presented a procedure for building a model that will provide reliable forecasts of the relevant uncertain quantity (the *Y*-variable), given knowledge of one or more influential factors (*X*-variables). For a model to be useful in this forecasting role, it must (1) make sense (i.e., be consistent with the model-builder's beliefs about the way things work); (2) be simple (i.e., not contain extraneous, insignificant terms); and (3) meet the assumptions underlying the model. Given that such a model has been built, exactly how do you use the model to forecast?

Point forecast

The first step in using the chosen model to forecast is an easy one. Simply substitute the relevant *X*-value(s)[20] into the fitted equation

$$\hat{Y} = \hat{a} + \hat{b}X$$

to calculate $\hat{Y}$, the point forecast of *Y* for the given value of *X*.

As an example, recall that earlier in this note we used 30 pairs of observations of new home construction cost, the *Y*-variable, and the corresponding square footage, the *X*-variable, to fit the model

$$\hat{Y} = \$56{,}104 + 63.11(X).$$

To forecast the construction cost of a proposed new 3,200-square-foot home, we simply substitute $X = 3{,}200$ into the least-squares equation given above to calculate $\hat{Y} = \$258{,}058$. Thus, the best guess for Lawrence Construction's cost of a 3,200-square-foot home, based on the fitted model and all the accompanying assumptions,[21] is $258,058.

If the decision only required a best guess for cost, our forecasting task would be complete. The single point forecast of $258,058 would suffice. Many decisions, however, demand an explicit consideration of the uncertainty surrounding a forecast. In particular, Jerry Baugher of Lawrence Construction wants to be sure that potential buyers understand just how uncertain the actual construction cost is for the proposed new home. To complete the forecasting task, then, we want a complete probability distribution for the cost of the proposed new home.

Interval forecast

The first step in obtaining the probability distribution for a new *Y* is to determine what shape this distribution will have. If we are willing to assume that the distribution of *Y* for a given value of *X* is normal, then the distribution of a new *Y* will also be normal.[22] Recall that a histogram of the residuals is a good way to check this assumption.

Once we decide to use the normal distribution, two quantities must be determined: the mean and the standard deviation. It makes sense that $\hat{Y}$, the point forecast, should be the mean of this distribution. Determining the standard deviation—the measure of how far *Y* is likely to be from $\hat{Y}$—is a little tricky.

[20] Throughout this section, we refer to the simple (one variable) linear model. The ideas presented, however, generalize to the multiple linear model.

[21] These assumptions (linearity, independence, homoskedasticity, and normality) are discussed in section 5.

[22] If *n*, the number of data points, is small, the t-distribution should be used instead of the normal distribution.

At first you might think that the standard error should be used. After all, the standard error is the estimate of the unexplainable, random error in *Y* for a given *X*. As such, it measures how far *Y* is likely to be from $a + bX$, the unknown mean of *Y*. If we did know the mean of *Y* (i.e., if we knew a and b and did not have to estimate them using our small sample of n observations), then the standard error could be used. But because we do not know the actual mean of *Y* and must rely on the imperfect estimate $\hat{Y}$, we face more uncertainty than is measured by the standard error.

When $\hat{Y}$ is used to forecast *Y*, there are two components of error that must be faced:

$$\begin{array}{ccccc} \text{Forecast error} & = & \text{Random error} & + & \text{Fitted error} \\ (Y - \hat{Y}) & = & (Y - \text{mean of } Y) & + & (\text{mean of Y} - \hat{Y}) \end{array}$$

The error in the forecast is composed of the random, unavoidable error in the process and the fitted error we make because our fitted regression line imperfectly estimates the true regression line.

Fortunately, theory provides a measure—the *standard error of fitted*—of the amount of fitted error faced. This standard error of fitted measures how far the actual mean of *Y* is likely to be from $\hat{Y}$. Memorizing the exact expression[23] for the standard error of fitted is not necessary, because this measure is included in many commercially available regression-software packages. It might be useful, however, to note that the standard error of fitted depends on three things: (1) the standard error (the amount of underlying uncertainty in the process dictates the amount of uncertainty in every estimate associated with the model); (2) the sample size (as expected, the standard error of fitted decreases with the square root of the sample size); and (3) the relative size of the known *X* used to forecast *Y* (the standard error of fitted is smallest for *X*-values close to the average of the *X*s and largest for extreme *X*-values).

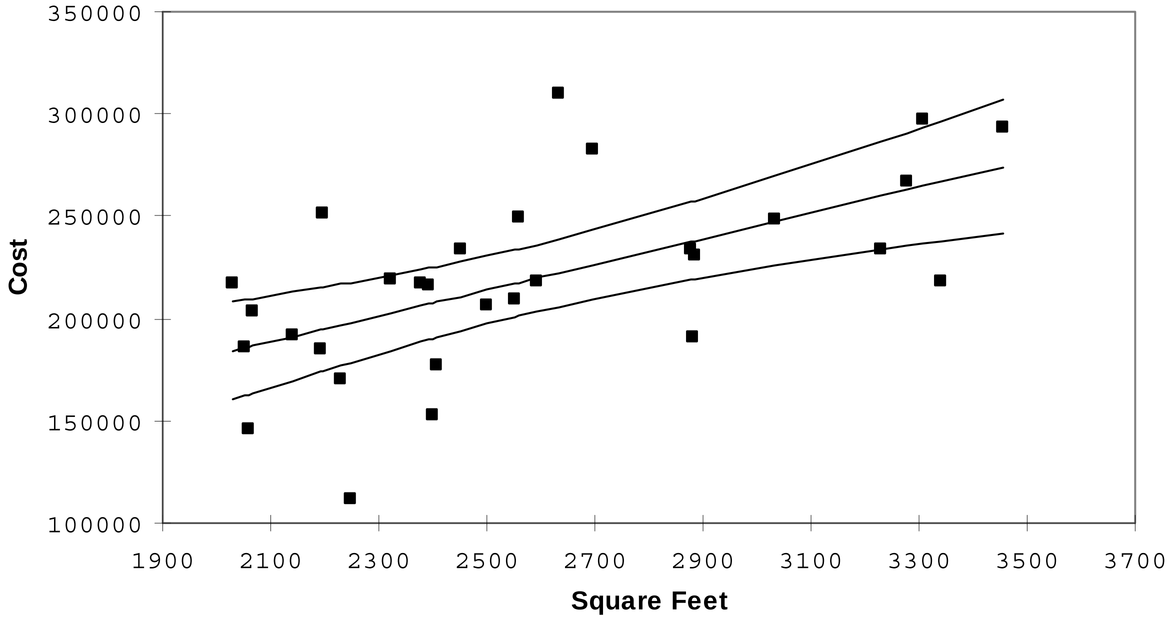
The standard error of fitted may be used to construct a confidence limit for the unknown, actual regression line. **Figure 13** graphs these confidence limits ($\hat{Y}$ plus and minus two standard error of fitted) for Lawrence Construction's cost as a function of the square footage of the new home. Notice that the confidence limits are smallest near the sample average of the *X*s and largest at the extremes. These curved confidence bounds should be intuitively appealing; they show the fitted model to be a better predictor at typical *X*-values than at extreme *X*-values.

Keep in mind, however, that these confidence limits only account for the sampling uncertainty in the fitted equation. They are curved because the effect of the uncertainty in *b*, the slope, is magnified at the extremes. A completely separate reason for a decrease in predictive ability at extreme *X*-values is that the model assumptions might hold only for a limited range of *X*-values. The possibility that the model assumptions are more tenuous at extreme *X*-values is not

[23] Standard error of fitted = Standard error $[1/n + (X - \overline{X})^2/\Sigma(X - \overline{X})^2]^{1/2}$, where *X* is the current *X* used to forecast *Y*, $\overline{X}$ is the sample average of the *n* observed *X*s, and $\Sigma(X - \overline{X})^2$ is the sum of squares of the *n* observed *X*s about $\overline{X}$.

accounted for in these confidence intervals. The limits are drawn assuming that the mean of $Y|X$ $= a + bX$ for all X.

Figure 13. Confidence intervals for the mean of $Y|X$.



Because the forecast error, or prediction error, is the sum of the random error and the fitted error, the uncertainty surrounding the forecast should be a combination of these two components. The measure of the uncertainty surrounding the prediction is called the **standard error of prediction**. The standard error of prediction is the appropriate standard deviation to use in the probability distribution for a new *Y*. Because the random error and the fitted error are independent and add together to form the prediction error, the standard error of prediction is the square root of the sum of the squares of the standard deviations of the components:

$$\text{Standard error of prediction} = [(\text{Standard error})^2 + (\text{Standard error of fitted})^2]^{1/2}$$

Many commercially available regression packages will also provide the standard error of prediction.

We have now completed the forecasting task. The distribution of a new *Y* will be normal, with mean $\hat{Y}$ and standard deviation equal to the standard error of prediction. Because the standard error of fitted is a component in the standard error of prediction, the standard error of prediction will be a function of all the things that affect standard error of fitted. In particular, as the sample size increases, the uncertainty in the fitted value decreases (standard error of fitted goes to zero), and the standard error of prediction will approach the standard error.

For the Lawrence Construction example, the standard error of prediction is calculated to be \$38,771, a value slightly higher than the standard error of \$36,879. The distribution of construction cost for the proposed 3,200 square foot home is normal (because $n = 30$ is small, it would be better to use the t-distribution with 28 degrees of freedom) with a mean of \$258,058 and a standard deviation equal to \$38,771.

Analogy to simple random sampling

This section has addressed several issues surrounding the use of a fitted simple linear model to forecast an uncertain Y given knowledge of a predictor X-variable. The concepts and terminology presented can be directly compared with simple random sampling, wherein a random sample of n observations from a normal population is used to forecast a new uncertain observation.

In simple random sampling, there is no predictor X-variable. The mean of Y is assumed to be a constant, and $\bar{Y}$, the sample average of the n observed Ys, is used as the point forecast of a new uncertain Y. The sample standard deviation, s, measures the uncertainty in Y about the mean of Y and is directly analogous to the standard error. The posterior standard deviation of the mean (calculated as $s/n^{1/2}$) measures how close the unknown mean of Y is likely to be to $\bar{Y}$. The standard error of fitted associated with the linear model is analogous to the posterior standard deviation of the mean.

When $\bar{Y}$ is used as the point forecast of a new Y, the relevant uncertainty is how close $\bar{Y}$ is likely to be to Y. The standard deviation that appropriately measures this uncertainty is a combination of s, the sample standard deviation of the population, and $s/n^{1/2}$, the posterior standard deviation of the mean. Again, these two standard deviations add in a squared way to give standard error of prediction = $s(1 + 1/n)^{1/2}$ as the measure of how close a new Y is likely to be to $\bar{Y}$. Notice that this standard error of prediction accounts for the two sources of uncertainty faced when $\bar{Y}$ is used to forecast a new Y: the sampling error associated with $\bar{Y}$ and the underlying uncertainty in Y. Notice also the similarity (both in calculation and in concept) between this standard error of prediction and the standard error of prediction associated with the linear-regression model.

Table 5 summarizes the comparisons between forecasting using the simple linear model and simple random sampling.

Table 5. Forecasting an uncertain Y-variable.

Using the Simple Linear Model	**Using Simple Random Sampling from a Normal Population**
Assumes: mean of $Y \mid X = a + bX$ independence homoskedasticity normality	Assumes random sampling from a normal population.
Samples *n* (*X*,*Y*) pairs.	Samples *n* *Y*s.
Point forecast is $\hat{Y} = \hat{a} + \hat{b}X$.	Point forecast is $\overline{Y}$.
The standard error measures the uncertainty in *Y* about the mean of *Y*.	The sample standard deviation, *s*, measures the uncertainty in *Y* about the mean of *Y*.
The standard error of fitted measures the uncertainty in the mean of *Y* about $\hat{Y}$.	The posterior standard deviation of the mean, $s/n^{1/2}$, measures the uncertainty in the mean of *Y* about $\overline{Y}$.
The standard error of prediction measures the uncertainty in *Y* about $\hat{Y}$.	The standard error of prediction, $s[1+1/n]^{1/2}$, measures the uncertainty in *Y* about $\overline{Y}$.

8. Using dummy variables to represent categorical variables

In a number of practical business situations, an important explanatory variable is categorical rather than numerical. For instance, an MBA's starting salary may be related to his or her undergraduate degree program (liberal arts, business, science, or engineering). The sales of a consumer packaged good may be related to its position in the store (check-out register versus regular store rack). Or the price-to-earnings ratio of a stock may be related to the presence (absence) of a dividend (irrespective of the size of the dividend). In each of these examples, the potential explanatory variable is not numerically scaled and thus not directly usable as an independent *X*-variable in a linear model. This section describes and illustrates a technique (called *dummy variables*) for transforming categorical information into numerically scaled variables suitable for use in a linear model.

Example

An extensive questionnaire was administered to several auto workers in four U.S. assembly plants. The variable of interest, *Y*, was a weighted combination of responses to several questions measuring job satisfaction. Management wanted to compare the levels of reported job satisfaction at the four plants. A small subsample of the data from two plants is given in **Table 6**.

In order to use a linear model to describe the relationship between job satisfaction and plant, we must first convert the plant information into a form suitable for use with a linear model.

We do this by creating a new variable, *D*, defined to equal 1 if the reporting employee is from plant B and 0 if the employee is from plant A. Variable *D* is called a dummy or 0–1 variable. We also stack the data into a single column as in **Table 7**.

Table 6. Reported Job Satisfaction for Two Plants.

	Plant A	Plant B
	59	67
	55	57
	61	57
	57	54
	55	56
	45	59
	64	38
	58	43
	48	52
	57	51
Average	55.9	53.4
St. Dev.	5.7	8.2

Table 7. The Stacked Job Satisfaction Data.

	D	*Y*
	0	59
	0	55
	0	61
	0	57
	0	55
	0	45
	0	64
	0	58
	0	48
	0	57
	1	67
	1	57
	1	57
	1	54
	1	56
	1	59
	1	38
	1	43
	1	52
	1	51
Average	0.50	54.65
St. Dev.	0.51	6.98

Figure 14 is the scatter plot of *Y* versus *D*. This scatter plot simply shows the two groups of satisfaction ratings. One must be careful in interpreting the chart because three pairs of points represent duplicate ratings (the rating of 57 for plant B; 55 and 57 for plant A). The duplicate points fall exactly on top of each other in the chart. Thus we only "see" nine points for Plant B and eight points for Plant A even though ten are in each group.

Let us now consider the linear model

$$\text{Mean}(Y) = a + bD.$$

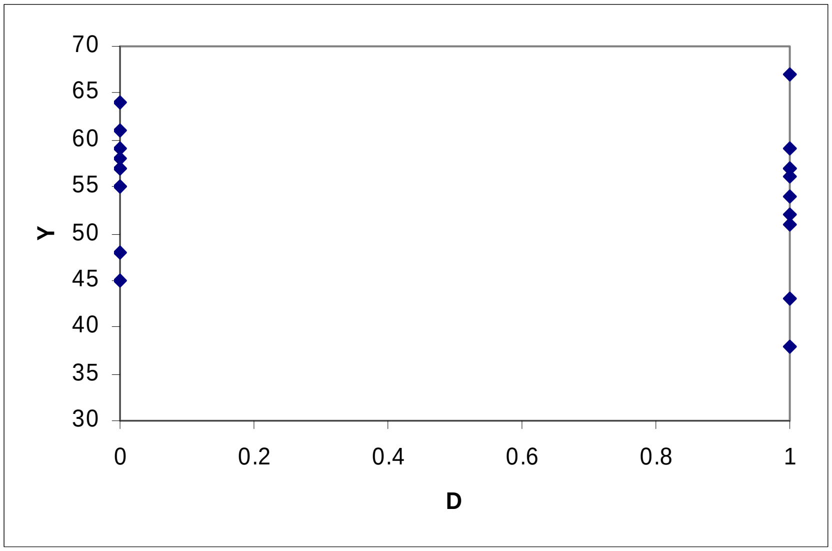
At $D = 0$ (corresponding to plant A), the model says that the mean of *Y* is equal to *a*, a constant. At $D = 1$ (corresponding to plant B), the model says the mean of *Y* is equal to a different constant, $a + b$. This model assumes only that the mean of *Y* is different for the two plants. When we fit the model to the data, coefficient *a* is the estimate of the mean of *Y* for plant

A, and coefficient *b* measures the difference between the mean of *Y* for plant B and the mean of *Y* for plant A. The model is capable of providing only two point forecasts, one corresponding to $D = 0$ (plant A) and the other corresponding to $D = 1$ (plant B). Not surprisingly, the fitted model turns out to be

$$\hat{Y} = 55.9 - 2.5\,(D).$$

This least-squares fitted line goes through the sample averages of the two groups of data.

Figure 14. *Y* versus *D* scatter plot.



Dummy variables for more than two groups

How would we use this dummy-variable technique if there were more than two groups? Suppose we had data from workers in all four plants. Could we define *D* as

$$D = \begin{cases} 0 \text{ if plant A} \\ 1 \text{ if Plant B} \\ 2 \text{ if plant C} \\ 3 \text{ if plant D} \end{cases}$$

and use the linear model mean(Y) = $a + b(D)$?

In general, the answer is no. This model, with only one *X*-variable, assumes that the mean of *Y* changes linearly as we move from plant A to B to C to D. It assumes that the means of the four plants fall on a straight line, and we usually have no reason to expect the group means to line up in such a manner. A less restrictive assumption is that the means of Y for the four plants can each be different, with no constraints on the ordering or relative size of the group means.

In order to incorporate this assumption into a linear model, we must define more than one dummy variable. Let D_B, D_C, D_D, be defined as

$$D_B = \begin{cases} 1 \text{ if plant B} \\ 0 \text{ if not} \end{cases}$$

$$D_C = \begin{cases} 1 \text{ if plant C} \\ 0 \text{ if not} \end{cases}$$

$$D_D = \begin{cases} 1 \text{ if plant D} \\ 0 \text{ if not} \end{cases}$$

and consider the multiple linear model

$$\text{Mean}(Y) = a + b_1(D_B) + b_2(D_C) + b_3(D_D).$$

Substituting, in turn, the dummy-variable values associated with each of the four plants shows that this model specifies a separate mean of Y for each of the four plants. The constant term in the model is interpreted as the mean of the baseline group designated if all dummies are set equal to 0 (plant A). Each of the b-coefficients is interpreted as the difference between the mean of the designated group and the mean of the base-line group.

Plant	D_B	D_C	D_D	**Mean(Y)**
A	0	0	0	a
B	1	0	0	$a + b_1$
C	0	1	0	$a + b_2$
D	0	0	1	$a + b_3$

In general, p-1 dummy variables are used to represent p groups or categories. Any of the p groups may be chosen to be the baseline group (designated if all dummies equal zero), and the remaining p-1 groups each correspond to a single 0–1 dummy. The linear model constructed in this manner assumes a separate and distinct mean for each group, with no restrictions on the ordering or relative size of the group means. Given the flexibility of a separate constant coefficient for each of the p groups, the least-squares fitted model will always go through the sample averages of all the groups.

9. Useful data transformations

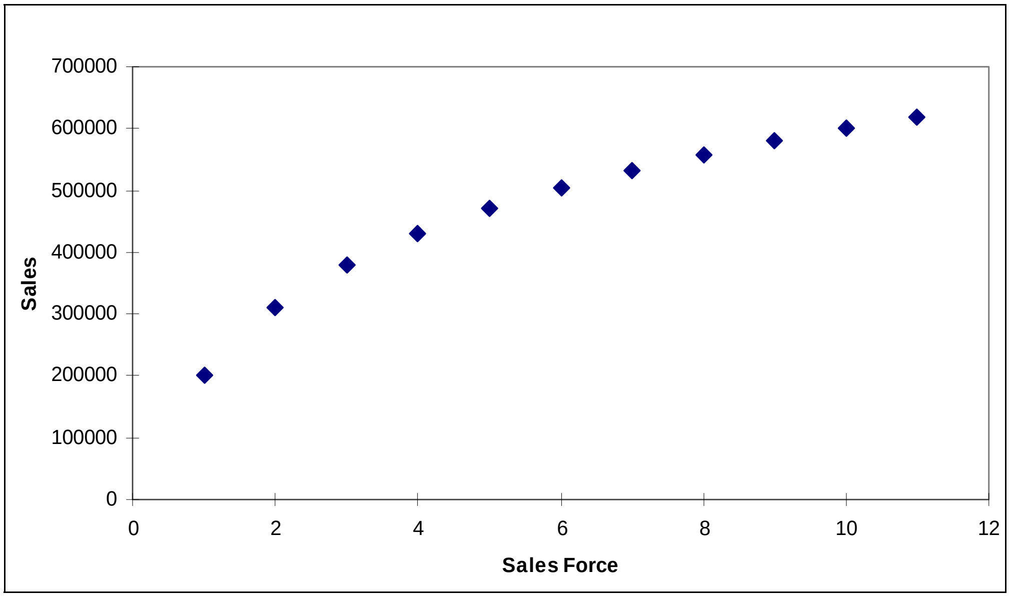
One of the important assumptions behind the linear model is that the mean of Y is linearly related to X:

$$\text{Mean of } Y \big| X = a + bX.$$

An implication of this assumption is that the change in the mean of Y for a unit increase in X is equal to b, a constant, regardless of the magnitude of X. For example, the point forecast of Y is b units higher at $X = 11$ than at $X = 10$, and it is also b units higher at $X = 91$ than at $X = 90$, even though on a percentage basis the change from 90 to 91 is much smaller than the change from 10 to 11. Although this linearity assumption is often reasonable, especially over a limited range of X and Y values, situations will arise where it is not tenable.

In some cases, the change in Y for a unit increase in X will decrease with the size of X. Think of the relationship between sales (measured in dollars) and the level of sales-force effort (measured in number of salespeople). As the number of salespeople assigned to a region increases, the resulting sales should increase, but at a rate that eventually decreases. Because customers will be serviced in decreasing order of their sales potential, assigning a larger sales force means going after increasingly smaller customers and thus decreasing the average sales per salesperson. A plot of sales versus sales force would show a nonlinear, *decelerating* relationship as in **Figure 15**.

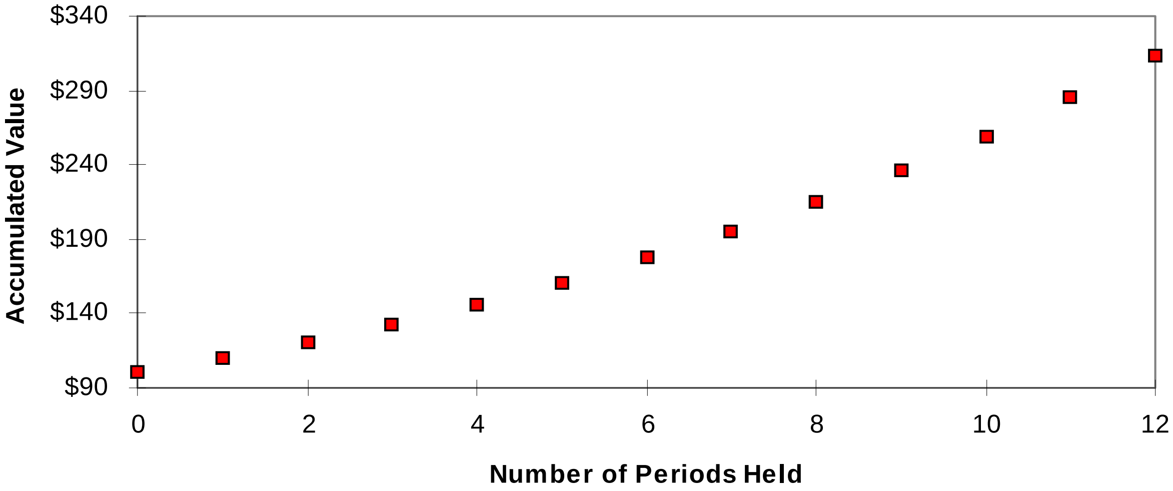
Figure 15. Decelerating relationship.



In still other situations, the relationship between Y and X may be nonlinear and *accelerating*. Think of the relationship between the accumulated value of an investment and the number of periods it is held. If the investment earns a fairly constant per-period rate of return over its life, its accumulated value will increase by an ever-increasing amount. A plot of accumulated value versus number of periods held (see **Figure 16**) shows this nonlinear, accelerating relationship.

This section addresses what can be done in those situations where it is not reasonable to assume that the relationship between Y and X is linear. The technique proposed is called transformation of variables, and the main idea is to carefully create a new variable from one of the old variables in such a way that the new, transformed variable can then be used in a linear model.

Figure 16. Accelerating relationship.



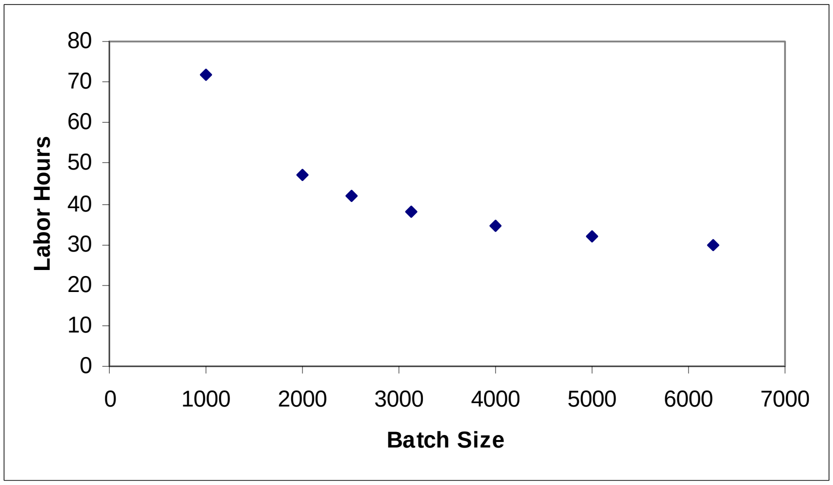
Example

For the last several months, a spare-parts manufacturing company has been conducting an experiment on the effect of batch size on productivity. They tried several different batch sizes and carefully kept track of the total direct labor hours required to produce their monthly production quota of 100,000 units. The results showed that total labor hours decreased with batch size.

A plot of hours versus batch size in **Figure 17** shows the relationship to be decidedly nonlinear and decelerating. The change in hours for a change in batch size from 1,000 to 2,000 (a decrease of 25 hours) is much larger than the change in number of hours for the same size change in batch size from 4,000 to 5,000 (a decrease of 2.5 hours).

Hours	Batch Size (number of units)
72.0	1,000
47.0	2,000
42.0	2,500
38.0	3,125
34.5	4,000
32.0	5,000
30.0	6,250

Figure 17. Total labor hours versus batch size.

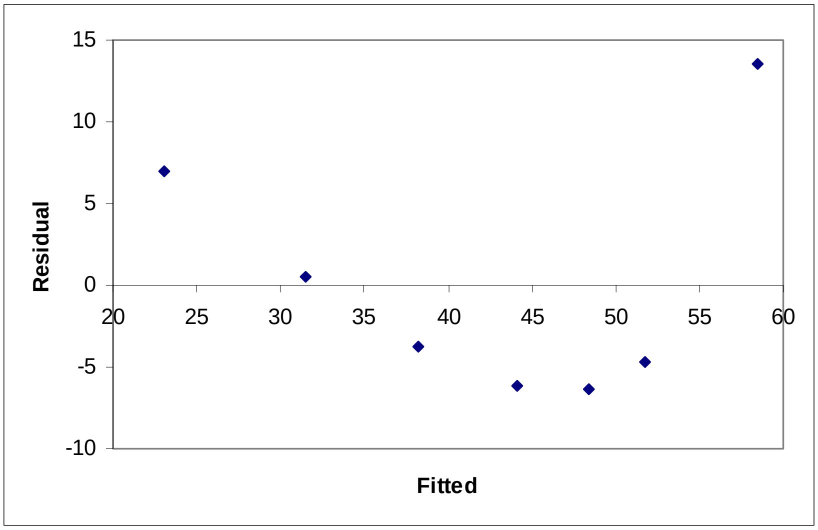


If we ignored this decelerating nonlinearity and fit the simple linear model

Hours = $a + b$ (batch size)

to these data, we might get a large R^2 and significant t-statistic. But a plot of the data or the model residuals (see **Figure 18**) would clearly show that this model does not fit. The model underforecasts at both high and low batch sizes and overforecasts at intermediate values. This systematic pattern in the residuals indicates that the model does not fit and should not be used.

Figure 18. Residual scatter plot showing a smile.



We now need some way to model this nonlinear, decelerating relationship between hours and batch size. One approach is to use a carefully chosen function of the batch-size variable as the independent variable in a new simple linear model. The creation of a new variable from an old one is called a *transformation*, and several ways are available to do it. The conversion of degrees Fahrenheit to degrees Celsius is one common example of the transformation of a variable. In the spare-parts manufacturer's case, we need to choose a transformation such that the relationship between hours and the transformed batch-size variable is linear.

The search for a useful transformation begins with an examination of the situation being modeled. If we ask ourselves why hours decrease with batch size, we realize that larger batches mean longer runs and fewer set-ups. Since a fixed number of parts (100,000) were made each month, the actual run time to produce the parts was fairly constant. Batch size will only affect the amount of time spent in set-up activities, and this set-up time is probably proportional to the number of batches produced. The number of batches can be calculated as 100,000 times the reciprocal of batch size, or

$$\text{Number of batches} = 100{,}000(1/\text{Batch size}).$$

This conversion from batch size to number of batches is an example of a transformation of variables. Because it makes sense that hours should be linearly related to number of batches, the simple linear model

$$\text{Hours} = a + b(\text{Number of batches})$$

will be tried. The coefficient a in this model represents the run time for the 100,000 parts, and the coefficient b represents the set-up time for each batch. Converting the available batch-size data to number of batches and plotting hours versus number of batches (see **Figure 19**) shows just how well this transformation works (and how carefully this example was constructed).

This transformation works because it changes the scale of the independent variable in just the right way. Notice that the change in number of batches for a change in batch size from 1,000 to 2,000 (a decrease of 50 batches) is much larger than the change in number of batches for the same size change in batch size from 4,000 to 5,000 (a decrease of only five batches).

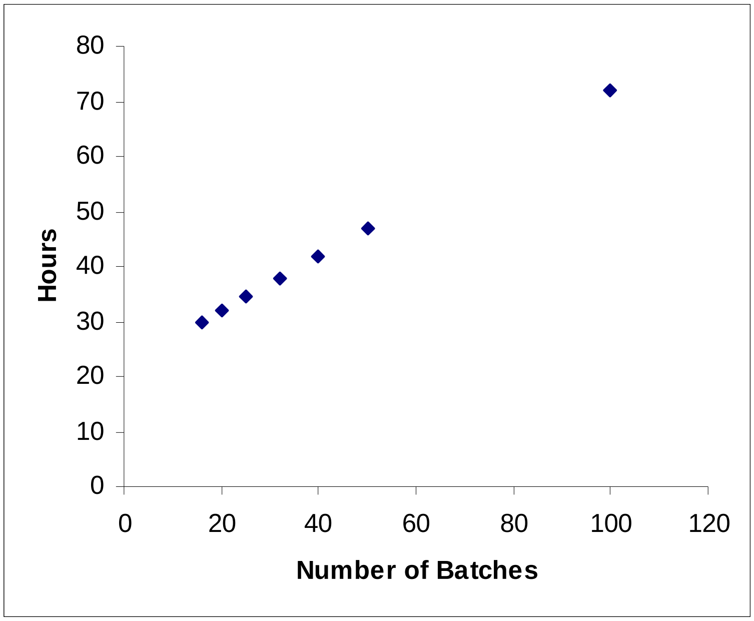
Hours	**Batch Size (number of units)**	**Number of Batches**
72.0	1,000	100
47.0	2,000	50
42.0	2,500	40
38.0	3,125	32
34.5	4,000	25
32.0	5,000	20
30.0	6,250	16

The fitted model

$$\text{Hours} = 22.0 + 0.5(\text{Number of batches})$$

fits the data perfectly, and shows that it takes 22 hours to make the 100,000 units and 0.5 hours to set up each batch.

Figure 19. Hours versus number of batches.



This simple example illustrates the use of a transformation of variables to model a nonlinear relationship between two variables. First it was noted that the relationship between hours and batch size was not linear, but rather decelerating. A plot of hours versus batch size confirmed this nonlinearity and suggested that the linear model would not fit. As an alternative, the model, Hours = a + b(Number of batches), was proposed. In one sense, this new model is linear; number of hours is expressed as a linear function of number of batches. The simple linear model and least squares can now be used with number of batches as the independent variable. In a separate sense, however, this model is nonlinear. If we rewrite it as

$$\text{Hours} = a + b(100{,}000/\text{Batch size}),$$

hours are expressed as a nonlinear function of batch size. The particular transformation chosen (100,000 times the reciprocal of batch size) was one that converts the nonlinear, decelerating relationship between hours and batch size to one that is linear.

Choosing a transformation

The criteria for choosing a transformation are identical to the criteria for building a model. Above all, the transformation should make sense (i.e., it must be consistent with the model-builder's belief about the underlying relationship between the variables). And second, the transformation should fit the available data.

In order to make an intelligent choice, then, it will be necessary to understand the implications of several candidate transformations and the relationships they represent. For example, in order to know that the reciprocal (1/*X*) is the appropriate transformation to use in a situation, you must know what a graph of *Y* versus *X* looks like if *Y* is actually a linear function of 1/*X*,

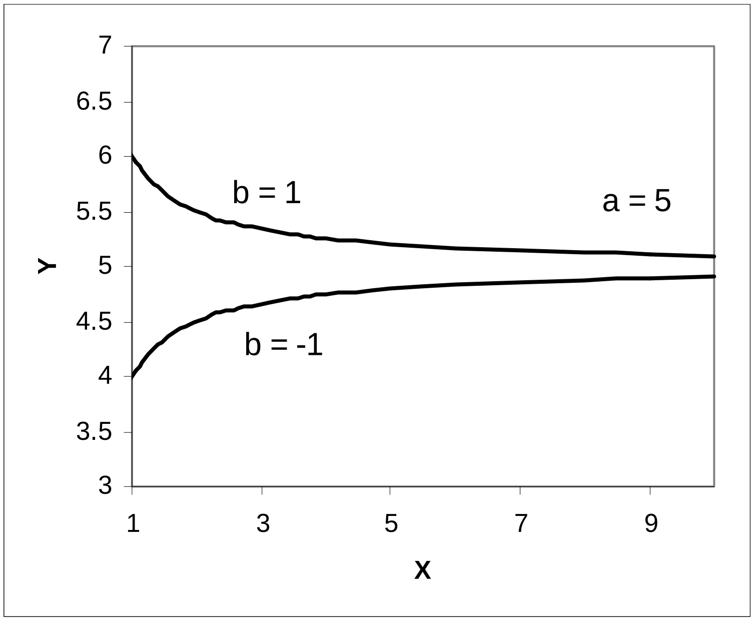
$$Y = a + b(1/X).$$

The remainder of this section presents and describes several useful transformations. Included in each section is a graph of what *Y* versus *X* can look like if that transformation is the appropriate one to use.

Reciprocal. The reciprocal of *X*, written as either 1/*X* or X^1, is defined as 1 divided by *X*. The hours/batch size example has already shown us that the reciprocal transformation can be useful when the relationship between *Y* and *X* is decelerating (see **Figure 20**).

Notice that because 1/*X* approaches 0 as *X* approaches infinity, coefficient *a* represents the limiting value of *Y* as *X* gets large. Because 1/*X* approaches infinity as *X* approaches 0, this transformation is not appropriate if *X* = 0 is a possible value. This transformation is sometimes used to represent the relationship between sales and price. At a price = 0, we expect a large value for sales, and as price increases, sales might decrease to some minimum level *a*.

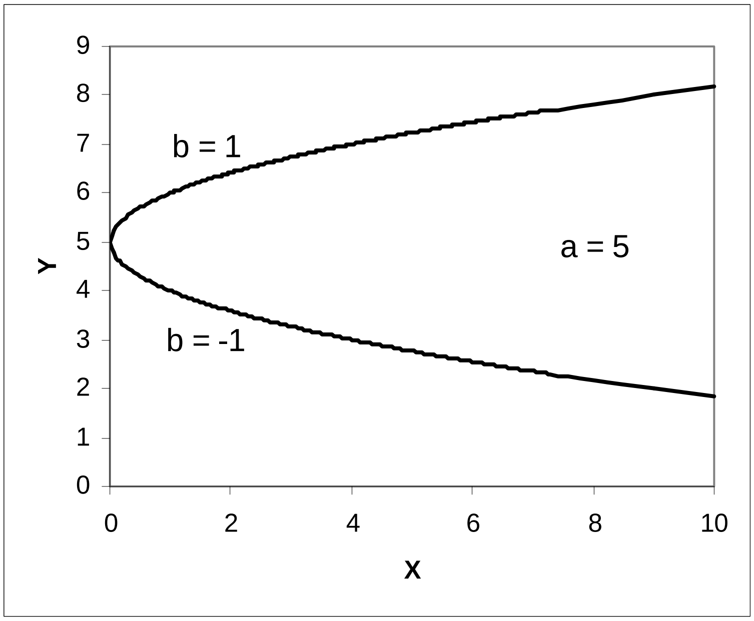
Figure 20. $Y = a + b(1/X)$.



Square Root. The square root of *X*, written as √*X* or $X^{½}$, is defined as the number that, when multiplied by itself, gives *X*. For example, the square root of 9 is 3, the square root of 4 is 2, and the square root of 0.25 is 0.5. (The square root is not defined for negative numbers and thus should not be used if *X* can be negative.) **Figure 21** graphs $Y = a + bX^{1/2}$ for both a positive and negative value of *b*.

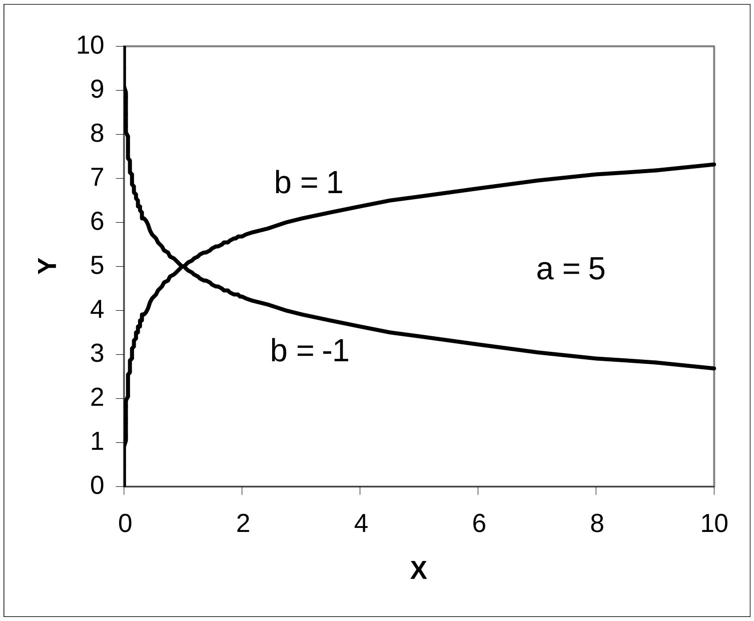
Notice that this relationship between *Y* and *X* is also decelerating, but at a slower rate of deceleration than that of the reciprocal. The square root is thus a less drastic transformation than the reciprocal and might be used if the relationship between *Y* and *X* exhibits only a small degree of nonlinearity.

Figure 21. $Y = a + b(X^{1/2})$.



Natural logarithm. The natural logarithm of *X*, written as *1n*(*X*), is defined as the number that, when used as an exponent of *e*, gives the value *X*. The value *e* is a carefully chosen constant equal to 2.718. Thus, the natural logarithm of 2.718 is 1, the natural logarithm of 1 is 0, and the natural logarithm approaches negative infinity as *X* approaches 0. (Like the square-root transformation, the natural logarithm is not defined for negative values of *X*.) **Figure 22** graphs $Y = a + b \times ln(X)$ for both a positive and negative value of *b*.

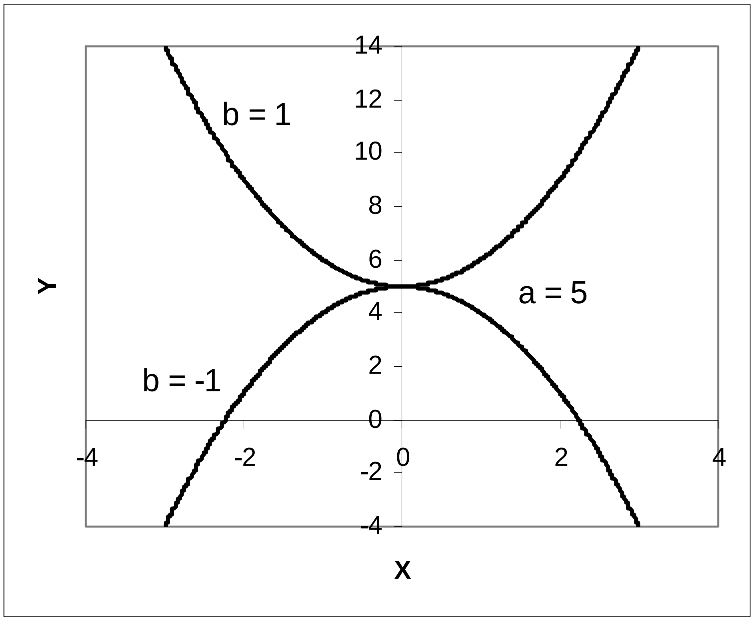
Figure 22. $Y = a + b \times ln(X)$.



The $1n(X)$ transformation also represents a decelerating, nonlinear relationship. It decelerates faster than the square root but not as fast as the reciprocal and is useful if the degree of nonlinearity is greater than that of the square root but less than that of the reciprocal. Because the natural logarithm of 1 is 0, the coefficient a represents the Y value at $X = 1$. As X approaches 0, $1n(X)$ will approach either plus or minus infinity (depending on the sign of b).

Square. A simple transformation that might be used to represent an accelerating relationship is the square. The square of X is defined as the product of X times itself. **Figure 23** graphs $Y = a + bX^2$ for both a positive and negative value of b.

Figure 23. $Y = a + b(X^2)$.



The square of X is defined for all values of X (positive and negative), and a graph of $Y = a + bX^2$ versus X is that of a parabola. The relationship graphed is called accelerating because the magnitude of change in Y for a unit increase in X *increases* with the magnitude of X (if X is positive). Raising X to powers greater than 2 also represents accelerating relationships, where the rate of acceleration is proportional to the size of the exponent used.

Transforming the Y-variable

So far we have discussed several possible transformations of the independent variable. It is also possible to use these transformations on the *Y*-variable. (Again, these transformations should be used if they help capture the underlying relationship between *X* and *Y*.) Transforming the *Y*, however, creates several complications.

Because the linear model is often used to forecast the dependent variable, transforming the *Y*-variable will change the nature of the quantity being forecast. Transformations of the Y-variable also affect the distribution of the model residuals, the interpretation of the measures of fit, and the pattern of scatter in the residuals (hetero- or homoskedasticity). These changes are sometimes desirable; one way to handle heteroskedasticity is to transform the *Y*-variable using the square root or natural logarithm. Keep in mind, however, that using a transformation on the *Y*-variable changes the forecasting problem; instead of forecasting *Y* and explaining a certain percentage of the variance of *Y*, you will now be forecasting the square root of *Y*, for example, and explaining a *different* percentage of the variance of the square root of *Y*.

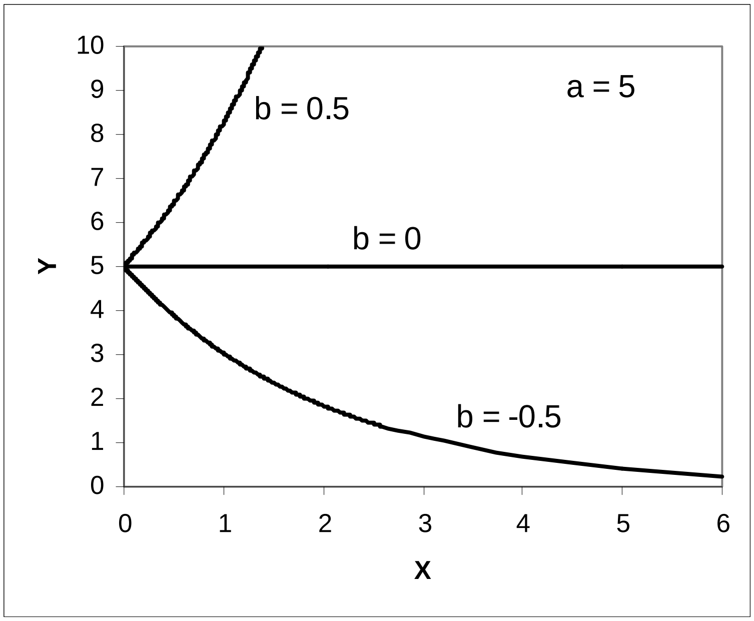
Exponential model—the learning curve. There are two important model forms that can be converted to a linear model with a transformation of the *Y*-variable. The first will be called the exponential model,

$$Y = ae^{bx},$$

in which *e* is the constant equal to 2.718. This model is used in business contexts (particularly in the field of operations), where it is called the learning curve (see **Figure 24**).

The model can represent either an accelerating ($b > 0$), decelerating ($b < 0$), or constant ($b = 0$) relationship between *X* and *Y*. To transform this model to a linear form, we take the natural logarithm of both sides. Because the natural logarithm of e^{bx} is, by definition, *bX*, we obtain[24]

Figure 24. $Y = a(e^{bx})$.



$$1n(Y) = 1n(a) + bX,$$

which is just the simple linear model with dependent variable 1n(*Y*) and independent variable *X*. The constant in this linear model, *1n*(*a*), is just the natural logarithm of *a*, the multiplicative constant in the exponential model.

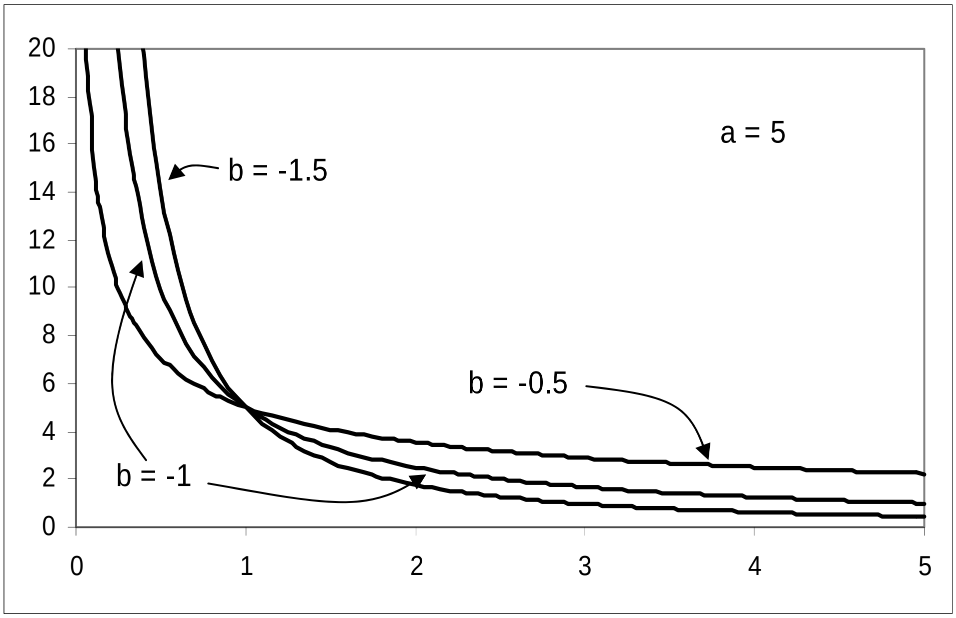
[24] We also use the property that *1n*(*UV*) = 1*n*(*U*) + 1*n*(*V*).

Multiplicative model—constant elasticity. The second important model form that can be transformed into a linear model is

$$Y = aX^b.$$

This model has the important property that the percentage change in Y for a given small percentage change in X is a constant for all values of X. The ratio of the percentage decrease in Y for an incremental percentage increase in X is sometimes called the elasticity of Y with respect to X. In the model, elasticity is a constant equal to $-b$. **Figure 25** graphs $Y = aX^b$ for elasticities of 0.5, 1, and 1.5.

Figure 25. $Y = a(X^b)$.



To transform this model to a linear form, we again take the natural logarithm of both sides in order to get

$$1n(Y) = 1n(a) + b \times 1n(X).$$

Thus, if $Y = aX^b$, the relationship between 1n(Y) and 1n(X) will be linear with intercept 1n(a) and slope b. The simple linear model can be used if both X and Y are transformed using the natural logarithm.

This model is called the multiplicative model, because the combined effects of several independent variables *multiply* together to determine the mean of Y:

$$\text{Mean}(Y) = a(X_1^{b1})(X_2^{b2})(X_3^{b3}).$$

This model is often used to represent the relationship between a product's share of market and marketing-mix variables such as price, advertising, and promotion. The multiplicative model is used if the effects of each of the marketing variables are assumed to depend on the levels of all the others. Taking the natural logarithm of this model,

$$1n(Y) = 1n(a) + b1 \times 1n(X_1) + b2 \times 1n(X_2) + b3 \times 1n(X_3)$$

converts it to a multiple linear model.