Harvard Business School 9-191-117
Rev. 2/18/93

Regression Analysis

When historical data are to be used in the assessment of a forecast, either the forecaster must believe the data come from indistinguishable situations, or adjustments must be made in the data to remove the effects of the causes which make the situations distinguishable from one another. Forecasting with regression analysis is essentially a process of identifying factors which bring about relevant distinguishability, evaluating their influences, and allowing for them in the development of a forecast.[1]

An Estimating Problem

To explore the reasoning involved in this process, consider the following estimating problem. A number of single-family dwellings located in a particular metropolitan area have been sold over the past year. A house will be selected at random from the list. Our task as a forecaster is to provide an estimate of its selling price. At this first stage no other information about the house, other than that it comes from the list, is provided. What information would we try to obtain before making a forecast? How would we use the information to make an estimate?

A moment's reflection will indicate the desirability of obtaining a complete list of the selling prices of the houses. If we were asked, along with others, to make an estimate of the prices of a randomly selected house from the list, the value we would select to be our estimate would depend upon the criterion by which the winner would be selected. If the winner would be any individual with an estimate exactly correct, the mode of the price list should be selected as our estimate. If the winner would be the individual with the smallest average error over a series of randomly selected houses, then the median of the price list would be the desired estimate.[2] If the criterion for judging the winning estimate would be the one with the smallest average **squared** error (over a series of randomly selected houses), then the arithmetic mean would be chosen.[3] In most situations in which regression is used, this last criterion, the so-called least squares criterion,

---

[1]There are other uses of regression analysis which raise issues not considered in this note.

[2]It can be demonstrated that the median is the value in a distribution from which the sum of the deviations, without regard to sign, is a minimum.

[3]The mean can be shown to be that value in a distribution from which the sum of the squared deviations is at a minimum.

This note was prepared as a basis for class discussion and is based on an earlier note "Developing Forecasts with the Aid of Regression Analysis" (175-105).

is used. In this note it will be assumed that the least squares criterion is being employed in judging an estimate.

In this estimation task, as we are asked to estimate the selling price for one house after another, we would make the same estimate for each house, namely the mean of the selling prices for all the houses on the list. At the end of our task we could then get a measure of how well we had done in our estimates by comparing our estimates to the actual selling prices and calculating the average squared error. The larger the average squared error, the poorer our set of estimates. The smaller the average squared error, the better our estimates would have been under our criterion.

As a second estimating task we will be asked again to estimate the selling price of a house selected from the list. This time, however, we will be given some additional information, namely, the size of the house. This house has 1,500 square feet of living area measured to the nearest 100 square feet. The entire list of sales no longer arises from occasions which are indistinguishable to us. Those situations in which a large house is sold are now **distinguishable** from those in which smaller houses were sold, so the entire list as it stands would not provide us with the best basis for the selling price of the house in question, nor would the mean of the selling prices provide us with a desired "best" estimate. If, however, we select from the total list, all sales in which the house sold was 1,500 square feet in size, these sales would contain data from situations which in our view are indistinguishable from the one we wish to forecast. The mean of this subset of prices would provide us with our estimate.

If we expected to be faced with making a series of such estimates for a number of houses of various sizes, we would classify the basic data on the list by house size and record the distribution of selling prices for each size. The average selling price in each size category would provide us with our estimates. We may find, for example, that houses with 1,500 square feet of living area sold on the average for $30,000, and those with 1,700 square feet sold on the average for $32,000. We would then take $30,000 as our estimate for a house with 1,500 square feet of living area, but $32,000 for our estimate for any house with 1,700 square feet.

After a series of such estimates has been made we can go back and check on how well we have done. If, in fact, selling price is related to the size of houses, then the average squared error in this later set of estimates should be smaller than those from our estimates developed when we had no other data. The stronger the relationship, the greater the reduction in the average squared error.

Our third estimating task will be again to estimate the selling price of a house selected at random from the list. This time we will have as information on which to base our estimate not only the size of the house but also its type of construction. All houses on the list have been classified as either (1) frame, (2) mixed frame and brick, or (3) all brick.

An extension of the preceding idea should suggest that we could simultaneously cross-classify all houses within the list by these two characteristics: size and type of construction. If the house for which we want to make the estimate is a brick house of 1,800 square feet, then the distribution of the selling price for all brick, 1,800-square-foot houses provides us with relevant data for assessing our sales price.

In concept this same procedure can be expanded to take into account other related pieces of information. If we were asked to estimate the selling price of a house from the list and were

told the size of the house, type of construction, and size of lot on which the house is located, a series of sub-lists of price—each conditional on a particular size of house, type of construction, and size of lot—would provide data for the appropriate estimate.

If data were free and unlimited in the amount available, estimating methods would never involve more than the use of the conditional averages and distributions described above. The successful application of the methodology would, however, require for most situations a huge amount of data—in some cases more than conceivably could be obtained; in others, more than could be justified economically. If we had only four factors which made situations appear distinguishable to us (four so-called explanatory variables), and if each of these factors could take on only ten values, we would have 10,000 cells (or categories) in our cross-classified table. If we needed, for example, 50 observations in each cell to get an approximation of the distribution of values in the cell, we would need a total of half a million observations to provide the basis for our estimates. If we want to add additional explanatory variables, the number expands rapidly. Thus, although this idea of basing estimates on sub-lists conditional on a given level of each of a number of explanatory variables provides us in concept with a powerful aid in estimating, in practice we rarely have, or could obtain, enough data to implement it directly. Regression analysis is a method for approximating what we would have deduced from the cross-classification analysis without demanding the voluminous amount of data that a direct application of that approach would have required. In such analysis a combination of relatively sparse data and some strong judgment on the part of the forecaster substitute for the extensive data otherwise needed.

The following example gives a rough idea of how judgment and sparse data can substitute for more extensive data. Imagine that in preparing for our estimates of selling price we have gone through the total list and selected a sample of 50 houses which were all 1,500 square feet in size. We calculated the average selling price and found it to be $30,000. We then took a second sample of 50 houses, all of which were 1,700 square feet in size. The average selling price of these houses we found to be $32,000. Before we have time to perform any additional analysis we are required to make an estimate for a house selected from the master list. We learn that the size of this house is 1,600 square feet (a size we have not specifically studied). What will we do? Although we have no direct data, most people's judgment would suggest that as the size of the houses increases, their average selling price would increase, and thus we would expect to find our estimate of the selling price for this house somewhere between $30,000 and $32,000.

Depending on the estimator's judgment of how selling price and size are related, a number of interpolations could be made. We might believe that from size group to size group of houses, each group differing from the previous one by 100 square feet, the average price increases by a constant amount (a linear relationship). Then we would make a linear interpolation and estimate $31,000 as the mean price, conditional on a size of 1,600 square feet as our estimate. On the other hand, if we believed there were some sort of diminishing returns in the relationship between selling price and size, we would interpolate a value greater than $31,000, for we would expect a greater difference between the average price of houses of 1,600 square feet compared to those of 1,500 square feet than between the 1,600-square-foot and 1,700-square-foot houses. A statement of the specific form of the relationship we believe to hold would allow us to obtain a specific value. Whether the value thus obtained would be very close to what we would have obtained had we had the extensive data depends primarily on how perceptive we were in understanding the nature of the relationships. If our perceptions correspond closely to what is actually going on, then our approximation will be close to what would have been obtained with extensive data, and the forecast will be a useful one. If our judgments are poor, we can still interpolate, but our forecast may be quite different from what would have been obtained from the extensive data and therefore misleading.

The arithmetic involved in developing regression estimates in any realistic problem is so extensive that it is almost essential that a computer be employed. Fortunately, almost every computer system has as part of its regular library a regression program. Programs may differ in the way in which information is to be provided and the exact form of the output, but all require the same kind of inputs from the user, and all provide in general the same information in the output. The next two sections will look at the kinds of information the computer requires of the user before it can perform regression analysis, and the kind of output the computer can usually provide to the user to convey the results of the analysis.

INPUTS TO A REGRESSION ANALYSIS

There are five kinds of information the user must supply to the computer before it can do its job. They are:

1. Identify the "dependent" variable. We have put quotation marks around the word "dependent" because it does not really refer to any characteristic of the variable itself. Rather, it is the customary way of identifying the variable that is being estimated or forecast. In the example above in which we attempted to estimate the selling price of a house based on its size and type of construction, selling price would be referred to as the dependent variable. With exactly the same data, if we wished to estimate the size of a house given knowledge of its selling price and type of construction, size would be the dependent variable. This requirement therefore states that if we wish to estimate or forecast some variable located in the computer data base, we must be able to indicate to the computer which variable it is we wish to estimate or forecast.

2. Specify the explanatory variable or variables. The explanatory variables[4] are those factors which, in the judgment of the forecaster, form the potential basis for distinguishability among the situations giving rise to the past data, and distinguishability from the situations for which estimates or forecasts are required. They are in essence the factors which it is hoped are related to the dependent variable and which will be used to "explain" differences among the values of the dependent variable. Selecting the explanatory variables to include in an estimating process is one of the important judgmental inputs required from the forecaster. The issues involved in these judgments are described in more detail in a subsequent section of this note, "Developing the Model."

3. Specify the relevant group of observations for the analysis. If the data base contains data on a set of observations larger than the group for which the analysis is desired, the forecaster must specify the subgroup of interest for the current analysis. If in the example above we were interested only in the selling price of single-family houses located in the metropolitan area, and if the total data base contained all property sales of that area, we might want to give the computer instructions to select for the analysis only that part of its total data base which refers to single-family houses.

4. Specify the nature of the relationship between the dependent variable and each of the explanatory variables. In the example of estimating the selling price for a house of 1,600 square feet, we saw that different estimates would result from different judgments regarding the nature of the relationship between size and price. A belief that the relationship was essentially a

[4]These variables are also referred to as "independent" variables.

linear one led to a estimate of $31,000. If we believed that the relationship was such that a difference in size of 100 square feet would have a greater effect on price if it were the difference between 1,000 square feet and 1,100 square feet than if it were between 2,000 square feet and 2,100 square feet (a retarding form of relationship), our estimate would have been greater than $31,000. If we state the specific form of the mathematical relationship, a unique value can be obtained. In a similar sense, if we believed that the relationship between size and price was one which accelerated (i.e., the average difference in price for the difference between 1,000 square feet and 1,100 square feet was **less** than the average difference in price between 2,000 square feet and 2,100 square feet), our estimate for the house of 1,600 square feet would be less than $31,000. Again, a statement of the specific mathematical relationship assumed would allow the calculation of a unique value. Before the computer can perform the calculation necessary to provide us with regression output, we must therefore specify a particular functional form of relationship between each explanatory variable and the dependent variable. The issues involved in deciding upon such relationships, the choices that are available, and the manner by which those choices can be communicated to the computer will be discussed in more detail in the subsequent section dealing with developing the model.

5. Provide data on the dependent variable and the various explanatory variables from all or a sample of observations from the relevant group. The basic goal of the regression process has been described as to combine judgments of the forecaster with relatively sparse data to obtain an approximation of results that would have required much more extensive data to obtain directly. This fifth requirement states that the appropriate relatively sparse data must be provided to the computer in the form of a data file. In the example in which we wished to estimate the selling price of houses based on their size and type of construction, a file containing the selling price, size, and type of construction for all or a sample of the single-family houses sold during the past year in the metropolitan area would have to be provided.

OUTPUTS FROM A REGRESSION ANALYSIS

If the forecaster provides the inputs mentioned above, the computer can do the arithmetic required for the regression analysis and provide the output. The various kinds of outputs from regression programs can be thought of as belonging to one of three categories: regression coefficients, measures of goodness of fit, and estimates or forecasts.

Regression coefficients

In the example of the estimate of selling price based on knowledge of the size of the house, if we were willing to specify that the relationship between price and size were linear, there are two numbers we would need to know to reproduce the entire cross-classified table of average selling prices for each size category. We would need, as a starting point, the average price in any one cell, and in addition, the constant change in the average price as the size of the house changed by a unit. If we were told that the average selling price of houses of 1,500 square feet was $30,000, and that for every 100-square-foot increase in size the average price increased by $1,000 (i.e., that in our judgment the relationship was linear between 1,000 square feet and 3,000 square feet), we could reproduce any entry in the table of average selling price conditional on size over that range. For example, we would estimate that the average selling price for houses of 2,000 square feet was $35,000 with this procedure: Starting with recognition that the average selling price for houses of 1,500 square feet was $30,000, and recognizing that a house of 2,000 square feet was 500 square feet

larger and that the average selling price increased by $1,000 for every additional 100 square feet of floor area, we would conclude that houses of 2,000 square feet would have an average price $5,000 greater than houses of 1,500 square feet or $35,000. Similar reasoning could allow us to estimate all of the conditional means of the table. The calculated values of the average change in the dependent variable per unit change in a given explanatory variable are known as **regression coefficients**.

Regression programs inevitably provide the user with a regression coefficient for each of the independent variables, as well as a starting point. Conventionally, the starting point is the average value for a dependent variable when each of the independent variables is set to zero. Thus, in the example the conventional starting point reported would be $15,000 (you can check to see that such a starting point would be consistent with a change of $1,000 per hundred square feet and a price of $30,000 for a house of 1,500 square feet). It should be noted that in many situations the starting point or constant does not have any interpretive significance. (In the example on selling prices, it does not make much sense to talk about houses with 0 square feet.) The starting point merely provides an agreed-upon and convenient value to which the changes can be added to get the estimate.

When there is only a single explanatory variable, a graphic depiction of the data and a geometric interpretation of the coefficients are possible.

Table 1

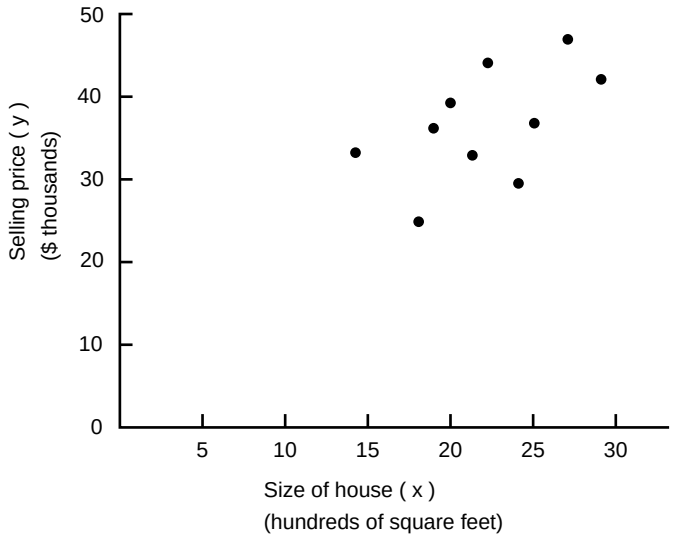
Selling Price and Size from a Sample of Ten Houses

Selling Price (y) *($ thousands)*	*Size of House (x)* *(hundreds of square feet)*
33.0	15
30.0	24
37.5	19
42.0	29
44.5	22
34.0	21
40.0	20
24.5	18
48.0	27
36.5	25

Suppose, for example, we wished to predict selling price of houses and we had foreknowledge of only the size of the house before we must make the prediction. Suppose also we had past data on the selling price and house size of a sample of ten houses, as shown in *Table 1* and graphed in *Figure 1.*

Figure 1

Scatter Diagram of Selling Price (y) and Size of House (x)



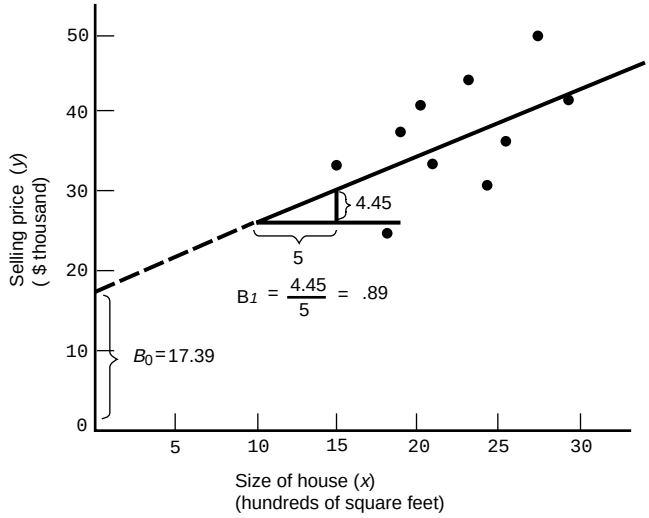
In *Figure 1*, called a scatter diagram, each dot represents the size of the particular house (x) and its sales price (y). The points show that the sales prices of the houses generally increased with increased size. Although the data are so sparse that we cannot calculate an average sales price for each size house, we might ask whether we believe the average relationship might fairly well be described by a straight line like the one shown in *Figure 2.* How this particular line was chosen will be described later. For the moment it is sufficient to note that such a line may be thought of as describing the average house value (y) for any house size (x) and thus it provides a reasonable forecast of the selling price of a house given any house size. This line represents a forecasting formula of the form

$$y_{est} = 17.39 + .8916x \ .$$

The slope of the line in *Figure 2* is the regression coefficient. If the line had been projected back to where the size of house was 0 square feet, it would have corresponded at that point to a selling price of $17,390. Thus geometrically the starting point corresponds to the point where the line intersects the y axis and the regression coefficient to the slope of the line when a linear relationship is specified.

Figure 2

Geometric Interpretation of Regression Coefficients



In the regression program we will use, and in fact in most regression programs, the constant and the regression coefficients are symbolized as B values (sometimes b, sometimes β). Thus, in an example with selling price as the dependent variable and with two explanatory variables—for example, size of house (in hundreds of square feet) and size of lot (in hundreds of square feet)—the computer output[5] might be of the following type:

y = Selling price
x_1 = Size of house in hundreds of square feet
x_2 = Size of lot in hundreds of square feet

B_0 = $10,000	Constant
B_1 = +$1,000	Change in y associated with a unit change (100 square feet) in x_1, the size of house
B_2 = +$500	Change in y associated with unit change (100 square feet) in x_2, the size of lot

The regression coefficients reported are those that, if used to estimate retrospectively in the data base, would provide a series of estimates which would have a smaller average squared error than any other comparable set of values. Although the coefficients are not actually calculated this way, we can think of it as following this procedure: In the example, with the two explanatory variables there are three coefficients (one constant and two slopes). The computer could pick three numbers, any three, for these coefficients and then use them to make estimates of selling price for all the observations contained in the analysis. When the estimates have been made they can be compared to the actual prices and the average squared error calculated and remembered. Then a second set of three numbers can be chosen for the coefficients and a similar procedure followed. If the computer swept through all possible sets of numbers, it could then report to us that set of values which provided the "best" set of estimates: the one with the smallest average squared error.

[5]This could be expressed as a linear equation of the form:

$y_{est} = B_0 + B_1x_1 + B_2x_2$.

In many situations the data used in the analysis comprise a sample of the complete set of data that could have been analyzed. In such cases the sample regression coefficients are approximations of the coefficients that would have been obtained had the complete set of data been analyzed. Most regression programs, in addition to reporting the regression coefficients themselves, also provide information which allows the user to evaluate how good an approximation each sample regression coefficient is of the coefficient that would have been obtained had the complete set of data, or at least a very large sample, been analyzed.

The measures of uncertainty for the regression coefficients are often labeled as **standard errors**. If a regression coefficient has value 1.2 (say) with standard error 0.5, then it is believed that if the regression were to be run with a much larger amount of data, then the corresponding regression coefficient for **that** output would be 1.2±0.5.

When the interest is solely in forecasting, the principal use of the regression coefficients is as a check on the sensibleness of the results. Potential explanatory variables are introduced in the analysis because the forecaster anticipates they will bear some relationship to the dependent variable and thus will increase predictive power and give a smaller average squared error. The reason for their inclusion will usually also suggest whether the relationship would be anticipated to be positive (+ sign for regression coefficient) or negative (- sign for the regression coefficient). Discovery of a sign or magnitude contrary to that anticipated should at least raise a red flag and encourage further investigation before the results are used to make forecasts.

You may develop, or be provided with, a very complex regression model and wish to determine whether the results make sense to you. In such cases, a simple interpretation of a coefficient and through it a simple test of "sensibleness" may not be possible. You can, however, make a series of forecasts in which you place all of the explanatory variables but the one under consideration at a fixed level and vary the value of the explanatory variable for which the reasonableness of the relationship is being considered. From the series of forecasts you can trace out the behavior of the forecasts as the explanatory variable takes on successively larger values and determine whether this pattern is a sensible one in the light of your knowledge of the factors involved.

MEASURES OF GOODNESS OF FIT

A second type of output provides measures of how good estimates using these regression coefficients were when the actuals and estimates were compared retrospectively in the data base. For any observation used in the regression the difference between the actual value of the dependent variable and its estimated value is called a ***residual***. The magnitude of these residuals gives the analyst some indication of the uncertainty that could be anticipated around new estimates or forecasts based on this regression. This type of information is reported in a number of different ways and using a number of different terms which largely reflect the preferences of the individual who wrote the particular computer program. In some cases, the **average squared residual** is reported and referred to by that name. In others, the same value is used, but it is called the **unexplained variance**. Because the units of the averaged squared residual are cumbersome, many people prefer to report the square root of this value, which is expressed in the original units of the dependent variable. The square root of the average squared residual is referred to as the **residual standard deviation**.

In the numerical example of the sales price and size of a sample of ten houses shown in *Table 1* above, a measure of goodness of fit can be calculated by the procedure shown in *Table 2.* The average squared residual, or unexplained variance, is shown to be 31. The residual standard deviation is the square root of this number, or $\sqrt{31}$ = 5.568 or $5,568.

Table 2

Calculation of Average Squared Residual (Unexplained Variance)

	Sales price (*y*)	*Size of house* (*x*)	*Estimated selling price** (y_{est})	*Residual* ($y-y_{est}$)	*Squared Residual* $(y-y_{est})^2$
	33.0	15	30.76	2.24	5.02
	30.0	24	38.78	-8.78	77.14
	37.5	19	34.33	3.18	10.08
	42.0	29	43.24	-1.24	1.54
	44.5	22	37.00	7.50	56.25
	34.0	21	36.11	-2.11	4.45
	40.0	20	35.22	4.78	22.88
	24.5	18	33.43	-8.93	79.81
	48.0	27	41.46	6.54	42.80
	36.5	25	39.67	-3.18	10.08
Average	37.0	22	37.00	0	31.00

* Based on the forecasting formula y_{est} = 17.39 + .8916x.

Whether we are considering the average squared residual or its square root, the interpretation is such that the larger the value, the poorer the set of estimates; the smaller the value, the better the set of estimates. If we were comparing two sets of estimates made for the same group of dependent variables, and if the basis for the two sets of estimates made equally good sense to us, we would probably prefer the estimating procedure with the smaller residual standard deviation. The interpretation of these measures is always a relative one. Is a residual standard deviation of $500 good or bad? If we have available an alternative procedure with a residual standard deviation of $50, then $500 is large, but if the next best forecasting procedure had a residual standard deviation of $1,000, then $500 would signal a considerably better forecasting process.

A second set of measures closely related to those described above is also customarily included in regression output. When the average squared residual for a set of estimates based on a regression is compared to the average squared residual from a set of estimates based only on the mean of the dependent variable (in the example above, the average selling price for all houses in a sample), a measure of the degree of improvement in the estimates accomplished through the regression is obtained. The customary comparison involves determining the percentage reduction in the average squared residual from the set of estimates found without regression (based solely on the average of the dependent variable) relative to the set of estimates determined by the regression process. For example, if the average squared residual in the set of estimates based on estimating selling price of every house at the mean selling price was 200,000 squared dollars and

the average squared residual in the set of estimates of the prices of the same group of houses based on a regression, taking into account a number of explanatory variables, was 60,000 squared dollars, you could indicate that there had been a 70 percent (or 140/200) reduction in the average squared residual. This percentage reduction, traditionally symbolized as R^2, is called the **coefficient of determination**, or the **percent variance explained**. Since this value is a percentage reduction, it varies between 0 and 100 percent. It is usually expressed as a decimal between 0 and 1. The greater the improvement brought about by the regression, the closer will this value be to 1; the less the improvement, the closer to 0.

Again referring to the numerical example involving the sales price and size of the house in the sample of ten houses, we listed from *Table 2* an average squared residual around the estimates of sales price based on the linear relationship with size of house of 31,000 squared dollars. If we had developed estimates of the value of each house **without** knowledge of size of house, we would have been forced to base our estimate on the average sales prices of the houses. *Table 3* shows the calculation of the average squared residual from such a forecast.

In this example, had we been forced to estimate solely on the basis of the average selling price, an average squared residual would have been 44,200 squared dollars.[6] Referring back to *Table 2*, recall that the average squared residual from a forecast made by taking into account the linear relationship between sales price and size was 31,000 squared dollars. The percent variance explained is the percentage reduction in the average squared residual, in this case, from 44.2 to 31.0. This percentage reduction becomes (44.2 - 31.0)/44.2 = 0.30. Thus in the example, 30 percent of the variance in selling price is "explained" by its linear relationship with the size of the house: $R^2 = 0.30$.

Table 3

Calculation of the Total Variance

	Selling price (y)	*selling price (Avg y)*	*Estimated Residual (y-Avg y)*	*Squared residual (y-Avg y)²*
	33.0	37	- 4.0	16.00
	30.0	37	- 7.0	49.00
	37.5	37	0.5	0.25
	42.0	37	5.0	25.00
	44.5	37	7.5	56.25
	34.0	37	- 3.0	9.00
	40.0	37	3.0	9.00
	24.5	37	-12.5	156.25
	48.0	37	11.0	121.00
	36.5	37	- 0.5	0.25
Average	37.0	37	0	44.20

[6]The average squared residual around the mean is, in this context, called the total variance.

Although this measure is on a standard scale from zero to one, its interpretation is still a comparative one. Does an $R^2 = 0.8$ indicate a regression we would want to employ? Not if we have an equally sensible process with an $R^2 = 0.9$. Is an analysis with an $R^2 = 0.2$ useful? Certainly, if the next best alternative for estimating has an $R^2 = 0.08$.

In most cases, the same comparative conclusions will be reached from any of the above sets of measures. The set of estimates with the larger average squared residual or residual standard deviation will usually be the set with the smaller percent variance explained.[7] If we wish to choose among any number of estimating methods that seem equally sensible to us on substantive grounds, the selection of that particular procedure with the smaller residual standard deviation, or, in the same sense, the one with the larger percent variance explained would allow us to pick the method that, at least in the past, would have done best.

THE ESTIMATES OR FORECASTS

Once the regression analysis has been run, we may wish to develop estimates of the dependent variable for new values of the independent variables. Such estimates constitute the third type of output provided by regression programs. In order to obtain an estimate from the regression we must provide a known or estimated value for each of the explanatory variables. The computer can do the arithmetic to provide the estimate of the mean value of the dependent variable, conditional on the stated level of each of the explanatory variables. In the example in which we wished to estimate the selling price of houses based on the size of the house (x_1) and size of the lot on which it is located (x_2), we had the following regression coefficients:

B_0 = \$12,000
B_1 = +\$10 per 100 square feet
B_2 = +\$0.50 per 100 square feet

If we now wish to have a forecast made, it is necessary to indicate the characteristics (in terms of the explanatory variables) of the house for which we wish to estimate the selling price. If we indicated interest in a house of 1,500 square feet located on a lot of 20,000 square feet, the computer would calculate the estimate by carrying out the following calculation:

Estimated selling price = \$12,000 + 10(1,500) + 0.50(20,000)
= \$37,000 .

Use of the resulting estimate as our forecast would still rely on our willingness to assume that the future situations for which the forecasts are required are indistinguishable from the situations that gave rise to the observation under analysis, in terms of their effect on the variations between the actual values of the dependent variable and the estimate from the regression. This would only hold true if, among other things, we were willing to assume that the forces and factors which gave rise to the relationship of the past would continue to operate in the same way in the future.

[7]This does not hold true necessarily if a transformation of the dependent variable is made for one of the sets of estimates. Such a situation will be described in the section on specification of the dependent variable below.

DEVELOPING THE MODEL

As we have seen, three important sets of specifications are required in preparing to use regression as an aid in forecasting. It is necessary to specify the dependent variable, to indicate the set of explanatory variables to be used as a basis for the forecasts, and to indicate the nature of the relationship between each of the explanatory variables and the dependent variable. The determination of these three inputs constitutes the construction of a regression model. This section deals with the issues and problems involved in making the required choices.

Selection of the Dependent Variable

We have noted that the dependent variable is the variable which is directly estimated by a regression model. It might appear, therefore, that knowing what we want to estimate or forecast automatically allows us to specify the dependent variable. In a sense this is true, and yet the scale and form in which we consider the dependent variable may have a major effect on the reasonableness of the assumptions underlying the analysis and thus on the usefulness of the results.

To return to the example in which we wanted to estimate the selling price of each of a group of houses, in addition to the selling price of each house in our data base, we might have information on size (square feet) and age (years). In the type of regression model we are considering here, an assumption is made that a given change in one explanatory variable adds a fixed amount to an estimate of the dependent variable when all of the others are held constant. It further assumes that this amount is the same regardless of the level at which the other explanatory variables are held constant.[8] If we were to use selling price as our dependent variable with size and age as two explanatory variables, the underlying assumption would be that a five-year difference in age, say, would add the same amount to our estimate of selling price for very large houses as for very small houses. We would almost certainly prefer an assumption that the average change in the price associated with a given age differential was greater for larger houses than for smaller houses. One way to accomplish this would be to specify our dependent variable not as sales price but as sales price **per square foot**.

In a similar way, if we were interested in estimating the sales of a particular product among various sales territories, we might have information on sales, population, median income, and percent urban population for each of the territories. If we were to select sales as our dependent variable, the assumption that a given difference in median income would bring about the same absolute difference in total sales in territories with large populations that it would in territories with small populations would almost certainly seem unreasonable. If we were to specify **per capita sales** as the dependent variable, the assumption that the effects of median income and percent urbanization are additive is much more reasonable.

We should be aware that if we convert the dependent variable from **selling price** to **selling price per square foot** or from **sales** to **sales per capita** it will have a major effect on some measures of goodness of fit. The coefficient of determination (R^2), for example, compares the

---

[8]There are other types of models possible (e.g., the multiplicative regression model). These go beyond the scope of this discussion, however.

average squared error around the regression estimates with the average squared error around the mean of the dependent variable. Since the scaling of the dependent variable in the manner decided above almost certainly reduces the total variance (the average squared error around the mean) more than the variance around the regression estimates, a much smaller R^2 would result. This is an inevitable consequence of the definition of R^2 and should not concern us in our choice of models.

If our real goal is to forecast **sales** but, in order to employ a model where the assumptions make more sense, we have used as a dependent variable **sales per capita**, we must convert the forecast made from the model to the form needed for the forecast. Developing a forecast of sales in a given territory from a model that generated a forecast of per capita sales requires only multiplication by the population of the territory.[9]

In thinking through the specification of the dependent variable, we ought to think carefully about the nature of the additivity assumption under the various alternatives for expressing the dependent variable. Selection of a form for the dependent variable which makes the additivity assumption most reasonable to us should provide a model in which we have greater confidence.

Selection of Explanatory Variables

If the number of observations available to us is very large relative to the number of potential explanatory variables, the selection is rather straightforward. We ought to include in our model all of the potential explanatory variables which we believe make sense in terms of our knowledge of the substantive area involved, and for which we will either know the value or have a good estimate of the value at the time we will be using the model to make a forecast. In some cases we may include variables which in our judgment do not themselves bear a relationship to the dependent variable but are related to an explanatory variable whose value is not known. In that context the included variable is brought into the analysis as a *proxy* for the unknown but related explanatory variable. When time is introduced as an explanatory variable, it is almost always in the role of a proxy variable. If we include an explanatory variable but it in fact does not improve our ability to forecast, nothing is lost as long as the number of observations is substantial.

Only variables which we will know or have a good estimate of at the time we need to make forecasts belong in our regression model. For example, it may well be true that historically the price of beef bears a close relationship to the price of pork. In a regression model designed to forecast the price of pork, it would not be useful to include the price of beef as an explanatory variable if we would not know beef prices, or at least have a better estimate of them than we could have of pork prices, at the time we had to make the forecast. It would not be sensible to include the original asking price of the house in forecasting the selling price even though they may be highly correlated if we are going to make forecasts of selling price before the original asking price has been established.

If the number of observations available for our analysis is not very large relative to the potential number of explanatory variables, a question of priority for including explanatory

[9]The standard deviation of the forecast of sales can also be determined by multiplying the standard deviation of the forecast for per capita sales by the given population.

variables becomes important. If two or more potential explanatory variables are themselves closely related, it would make sense to include only one in the restricted group of explanatory variables. At the extreme, if we had two measures of the size of a house, one in square feet and the other in square meters, although either one may be closely related to selling price, it would not make sense to include both. In a less extreme example, but based on the same reasoning, if because of sparse data we were limited in the number of explanatory variables, we would not want to include both the per capita disposable income of the area in which the house is located and the median family income of the area. Although they are not identical measures, they are so closely related that most of the predicting power of one is contained in the other.[10] We would want to select the one that we thought made more sense and eliminate the other from our model. If the number of remaining potential explanatory variables, all of which have a sensible claim for inclusion, is still large, we may wish to select for inclusion in our forecasting model the subset that produces the lowest estimated residual standard deviation.

Lagged Variables

When the data available for the regression analysis are in the form of a time series, where observations refer to different points in historical time, we may wish to use information from one period as an explanatory variable to help us forecast the value of the dependent variable in another period. If we were making an annual sales forecast for a given product for 1975, we might use the 1974 sales as an important explanatory variable. We can think of the original data base as a table, the columns of which refer to the variables and the rows to the observations. *Table 4* shows a part of a data base in which the annual sales and other characteristics of a given product are recorded.

Table 4

Year	*Sales (thousands of units)*	*Advertising expenses ($ thousands)*	*Average price ($)*
1970	872	250	6.10
1971	915	275	6.25
1972	950	290	6.30
1973	1,020	320	6.30

If we wished to include the previous year's sales as an explanatory variable in our model, we would have to instruct the computer to construct a fifth column in the data base. In the fifth column would be entered the sales of the previous year. Thus, in this column opposite 1971 we would find the sales of 1970; opposite 1972, the sales of 1971, and so on. *Table 5* shows a segment of this expanded data base derived from *Table 4*. The newly created variable, previous year's sales, is one example of a ***transformed*** variable. In this case, the transformation is a ***lag***; in the next section we shall discuss transformations that are dummy variables.

[10] In technical terms, the two variables are said to be collinear.

Table 5

Year	*Sales (thousands of units)*	*Advertising expenses ($ thousands)*	*Average price ($)*	*Previous year sales (thousands of units)*
1970	872	250	6.10	-
1971	915	275	6.25	872
1972	950	290	6.30	915
1973	1,020	320	6.30	950

Dummy Variables

In some situations a potential explanatory variable we wish to include in our model may not be measured in units like dollars, square feet or years but instead may describe a qualitative category into which an observation can be classified. In the example of developing a procedure to estimate the selling price of houses, the type of construction (frame, mixed, or brick) illustrates such a situation. We do not say that a house is some number (such as 1.75) on a "construction **scale**": rather we think of a house as falling exactly into one of the three **categories**. This type of explanatory factor can be incorporated into our model by the establishment of a "dummy variable" system. A dummy variable system requires the addition to the data base of a number of variables which is one less than the number of categories in the classification. In our example with the three categories (frame, mixed, or brick) two new variables would have to be established. Values for a variable in a dummy variable system are either 0 or 1. The 0 indicates that the observation does not possess the characteristic described by the particular variable. The value 1 indicates that the observation does have the characteristic described by the variable.

In introducing the type of construction into our model to predict selling price, since there are three categories (frame, mixed, or brick) we would add two dummy variables to our data base. The first dummy variable could refer to frame construction. In the data base the number 1 would be put under that variable for every house of frame construction. A zero for that variable would indicate that the particular house was **not** of frame construction (and was either mixed or brick). A second additional variable might refer to mixed construction. Observations relating to houses with mixed construction would have a 1 recorded for this variable. A house that was either all frame or all brick would have a 0 recorded under this variable.

To put a third variable in this system relating to brick construction would be redundant. Any house that had values of 0 under both the frame and the mixed variables **must** have a 1 under brick. Any house that had a 1 under either frame or mixed variables **must** have a 0 for a brick construction variable.

When a dummy variable is used, the resulting regression coefficients indicate the difference between the average value of the dependent variable for the category identified by that variable and the average value of the dependent variable for the category not explicitly included in the system. In our example, if the regression coefficient relating to the dummy variable called frame was - $3,000, it would indicate that, on the average, frame houses sold for $3,000 less than brick houses (the category not explicitly included in the system). Had we arbitrarily chosen to form a variable for brick but not explicitly included one for frame, we would have obtained a regression coefficient for brick of + $3,000 to show that brick houses sold for $3,000 more than frame (the excluded category). Similarly, in a system where we chose to eliminate brick from explicit inclusion, we might have obtained a regression coefficient for the mixed variable of - $1,000. This would indicate that for two houses identical in all other characteristics included in

our model, but where one was brick and the other of mixed construction, we would have an estimate for the latter that was $1,000 lower than for the brick house.

To review: the procedure to follow to include qualitative classifications as an explanatory factor in a model involves establishing **one less dummy variable** than the number of categories in the classification. Under a given variable the value of 0 is given if the observation does *not* have the indicated characteristic, and the value of 1 if it does. The resulting regression coefficients will then indicate the difference in the average value of the dependent variable from those observations that have the characteristic **not** explicitly included as a variable in the system.

Determining the Nature of Relationships

Once we have specified the set of explanatory variables we wish to include in our model, we must further specify the way in which we believe each of these variables relates to the dependent variables. The decision regarding the nature of a particular relationship springs largely from our knowledge of the substantive area in which we are forecasting. Our choice of the type of relationship we wish to specify between the selling price and size of house would come largely from our perception of what goes on in the real estate market. If we think of a group of houses that are the same in all characteristics considered by our model other than size of house, we might ask ourselves questions about what we would anticipate in terms of differences in the average selling price which will accompany differences in size.

If we believe that when houses differ by 100 square feet of floor space their average selling price will differ by a given amount, and that this amount will be the same regardless of where in the range of possible sizes we find the two groups of houses (e.g., 1,000 v. 1,100, or 2,000 v. 2,100), then we should specify a linear relationship between size and average selling price in our model.

In contrast, if we believe that a given difference in size (e.g., 100 square feet) would be accompanied by a larger difference in price if the difference were between 1,000 and 1,100 square feet than if it were between 2,000 and 2,100 square feet, we would specify some form of curvilinear relationship, where we would expect to find decreasing returns to scale. On the other hand, if we felt, based on our knowledge of the real estate market, that a given difference in size would, on the average, be accompanied by a larger difference in price for large-size homes compared to small-size homes, then we would want to specify some form of curvilinear relationship that would reveal increasing returns to scale.

A useful transformation of variables that makes it easy to conduct exploratory analysis to see if such curvilinear relationships exist in the data is described in the *Appendix*.

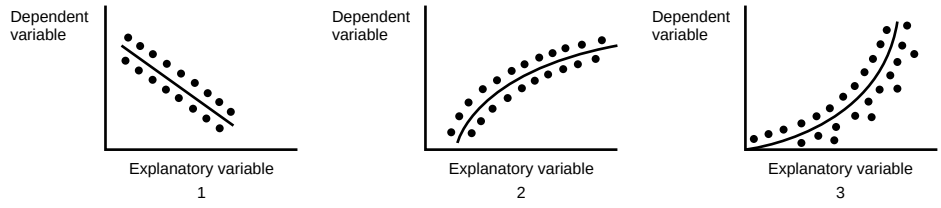
Two Warnings in Developing Models

There is a natural tendency to look for help in making the often difficult choices involved in specifying the nature of each relationship. However, there are difficulties to be encountered in looking at the data in order to get such help.

The fishing expedition. One suggestion that might come to mind, particularly since a computer is available to do all of the calculation involved, is to try out all the possible explanatory variables in all possible combinations of forms. If we have five potential explanatory variables and, let us say, seven different transformations of each variable which could be used to express each relationship, we could run the 16,807 different regressions and select, for example, the one with the highest percent variance explained.[11] The form of each relationship in this "best-fitting" model would constitute the forms we would then use in our final model. There are a number of studies that show that such a procedure runs a grave risk of suggesting a model which does not make sense and which does **not** have predictive power. A common experiment sets up a dependent variable and a number of explanatory variables, all from a table of random numbers. Regressions are run for all combinations of the explanatory variables and transformations of such variables. Inevitably, by the vagaries of chance, some peculiar combination just happens to give a set of estimates close to the values in the data base and a large value for R^2 results. How much use would that model be to us in forecasting the next value of the dependent variable? Since the next value of the dependent variable is determined from the next random number, none. The explanatory variables and their relationships must make sense to us in the light of our knowledge of the area in which we are forecasting before we would be willing to use the model to help provide our assessment of the uncertain future.

The two-dimensional graph. Another common tendency, in trying to decide on the nature of the relationship between a given explanatory variable and the dependent variable, is to plot a graph in which the explanatory variable is shown on the horizontal axis and the dependent variable is shown on the vertical axis. For each observation a point is made at the intersection of its value on each scale. The shape of the cluster of points on such a graph (called a scatter diagram) could then be used to suggest the nature of that relationship. The scatter diagrams shown in *Figure 3* indicate this procedure.

Figure 3



Observing scatter diagram 1 would lead us to specify a linear relationship. Having observed scatter diagram 2, we would probably specify a curvilinear relationship with decreasing returns to scale. Scatter diagram 3 would suggest a curvilinear relationship with increasing returns to scale.

There is a danger in following this procedure too literally. The nature of the relationship we are required to specify is that of a **net** relationship between the explanatory and dependent

[11] Since each of the five variables could be specified with each of the seven forms, the total number is $(7)^5$.

variables. The term "net" in this context means that all other explanatory variables contained explicitly in our model are held constant. In our example relating to the selling price of homes we would have to specify the nature of the relationship between selling price and size of home for houses of the same type of construction, located on lots of the same size, and so on.

The relationship that is observed in direct plotting of an explanatory and dependent variable is in contrast a **gross** relationship. In this context the term "gross" indicates that the relationship is depicted while the other explanatory variables are not controlled and can take on a variety of values. If the explanatory variables are not only each related to the dependent variable but related to one another, it is possible that the gross relationship might be of a quite different nature than the net relationship which is desired. A rather extreme example may illustrate this point. Assume we know for certain that there is no net relationship between the selling price of a house and the number of windows in the house. That is, for houses of exactly the same size, type of construction, and type of lot, the average selling price is the same for varying number of windows. If we were to plot a scatter diagram of selling price against the number of windows we would find a direct relationship in the diagram, since houses with a small number of windows typically are small houses, and the smallness of the house is, in general, associated with a low selling price. Large houses, on the other hand, have a large number of windows in general, and thus through the mechanism of the size relationship, the number of windows is positively related to the selling price. If, as in this example, a scatter diagram of a gross relationship could show a strong positive relationship when we know there was no net relationship, it should also be clear that the gross relationship could look like one shape while the net relationship was of some other form.

The results from plotting two-variable gross scatter diagrams may be suggestive but should not be followed too strongly. If the shape suggested does not make sense in the context of the problem, it would be unwise to use that relationship merely because of the scatter diagram. On the other hand, if we believed that the particular explanatory variable was essentially independent of the other explanatory variables, the plot of the scatter diagram may stimulate our thinking and improve our understanding of the phenomena we are attempting to model.

Protection against overfitting. It is, of course, difficult and in fact undesirable not to have some aspects of the data affect our choice of model. The patterns we observe as we run preliminary analyses with our data often suggest some "sensible" forms of relationship. Preliminary results frequently suggest what variables we wish to include and exclude from our final model. If we follow this natural procedure there is a danger we may be **overfitting**; that is, we may be selecting a model that matches the idiosyncrasies of the **particular sample of observations** under analysis but does not perform in nearly the same way when applied to new sets of observations from the same population.

If we have an adequate amount of data, one way to guard against overfitting is to develop the model and determine the regression coefficients from one part of the data and then apply the results to obtain forecasts from the other part. We can divide, in an appropriate way, all of the observations into two sets. The first set can be used to determine a preliminary model and its coefficients. The second set of observations can be used to determine the results obtained when using that model to make estimates for a new set of observations.

If overfitting has taken place and the preliminary model takes into account primarily the peculiar characteristics of the particular sample of observations from which it has been developed, this should show up in a comparison of the estimating errors from the two sets of observations. The fitting procedure guarantees that the average error for the set of observations used to determine the coefficients will equal zero. We can now use that same model to make an estimate of the dependent variable for each observation in the second set (the set not involved in developing the model or determining the coefficients). If the average error from this set of

observations is substantially different from zero, it would show an undesirable bias in the model when used in estimating new observations. If a comparable measure of the average **squared** error is substantially larger from the second set than from the set that suggested the model and provided the regression coefficients, we would know that overfitting had occurred.

If evidence of overfitting occurs either through a nonzero average error or a poorer goodness of fit, or both, the preliminary model should be viewed with skepticism. Our goal is to develop a model which has the best fit of those models that give consistent results when applied to the second set of observations. Once this model has been found, the final coefficients can be determined by fitting the model to the combined set of observations.

The key to developing a model useful to us in making forecasts for uncertain quantities lies in our perception of the factors that relate to changes in the variable that are being forecast. A regression analysis merely develops and quantifies the implication of the assumptions and judgments we introduce into the model. In the last analysis, the usefulness of the forecasts that result will depend not on the mechanics of the regression but on the soundness of our assumptions and judgments.

Appendix

Detecting, Specifying, and Interpreting Curvilinear Relationships: Exploratory Analysis

In regression models with several explanatory variables, we may believe that the "net" relationship between some one of the explanatory variables (let's call it x) and the dependent variable (y) is curvilinear, rather than linear. It is natural, under these circumstances, to plot a scatter diagram of x against y to see whether such curvilinearity is apparent. We have been warned, however, that such a diagram would show only the "gross" relationship between y and x ; the net relationship might be quite different.

A better way to detect curvilinearity under these circumstances would be to run a regression with x and the other explanatory variables included, compute the residuals, and plot the residuals (on the vertical axis) against x (on the horizontal). If this plot looks curvilinear, it suggests that the ***net*** relationship between y and x is curvilinear.

An even easier "quick and dirty" method of detecting curvilinearity is to include both x and a squared transformation of x (i.e., x^2) in the regression model. Holding the values of the other explanatory variables constant, the relationship between x and the estimated value of y is given by

$$y_{est} = b_0^* + b_1 x + b_2 x^2 \qquad (1)$$

where b_0^* is the sum of the constant term plus the products of the regression coefficients for the other explanatory variables times the values at which those variables are held constant; b_1 and b_2 are just the regression coefficients for x and for x^2.

For given values b_0^*, b_1 , and b_2 , a graph of y_{est} as a function of x will be a ***parabola***, a curve that either rises to a peak and then descends, or that descends to a trough and then rises. A value of b_2 clearly different from 0 provides evidence of a curvilinear net relationship between x and y ; if, on the other hand, b_2 differs from 0 only by chance (sampling error), the data do not supply strong evidence for a curvilinear relationship.

The addition of a squared transformation of an explanatory variable thus provides an easy way of ***detecting*** curvilinearity, but ***understanding*** in what way the relationship between y and x is curvilinear is a trickier matter. To interpret the nature of the relationship correctly, two facts about parabolas are important to know:

1. If b_2 (the regression coefficient for x^2 in equation (1)) is ***negative***, the parabola rises to a peak and then descends, while if b_2 is ***positive***, the parabola first descends to a trough and then rises again;

2. The trough or peak occurs at the value of x for which

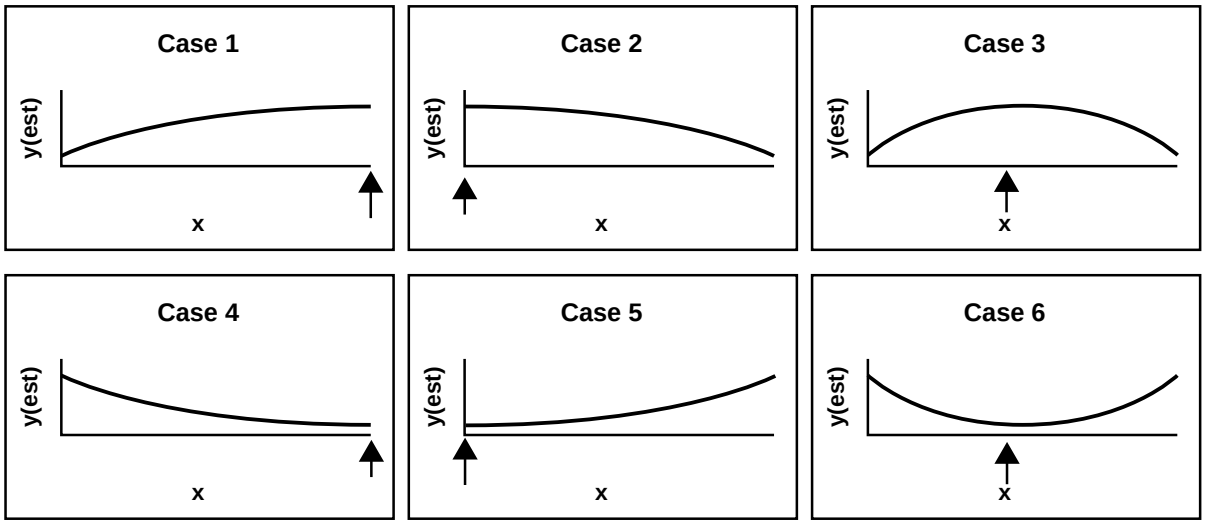
$$x = -b_1 / (2b_2) \quad .$$

Various forms of behavior depend on whether b_2 is positive or negative, and whether all, or nearly all, of the values of x are on one side or the other of the critical value $-b_1/(2b_2)$, or whether they substantially straddle that value. There are six cases to consider, as follows:

Case #	*Values of x*	*Value of b_2*	*Behavior of y_{est}*
1	$x < -b_1/(2b_2)$	$b_2 < 0$	Increases at a decreasing rate
2	$x > -b_1/(2b_2)$	$b_2 < 0$	Decreases at an increasing rate
3	x straddles $-b_1/(2b_2)$	$b_2 < 0$	Increases to max., then decreases
4	$x < -b_1/(2b_2)$	$b_2 > 0$	Decreases at a decreasing rate
5	$x > -b_1/(2b_2)$	$b_2 > 0$	Increases at an increasing rate
6	x straddles $-b_1/(2b_2)$	$b_2 > 0$	Decreases to min., then increases

Figure 4 shows these cases. The arrows indicate the location of $-b_1/(2b_2)$ relative to the values of x in the data.

Figure 4



If, for example, $b_1 = 17.43$ and $b_2 = -2.367$, then $-b_1/(2b_2) = 3.682$ and if most of the values of x are above 3.682, *Case 2* applies: y_{est} decreases at an increasing rate as x increases.

This method of analysis, it should be repeated, is ***exploratory***: it will usually reveal curvilinear relationships, and permit you to classify them appropriately. Three points are worth making, however:

1. When introducing a squared transformation (x^2), be sure to include the original value of x in the model as well;

2. The method discussed here will not reveal more complicated curvilinear behavior—for example, a relationship between y and x in which y first increases at an increasing rate and then at a decreasing rate as x increases;

3. Even if the general curvilinear relationship between y and x is detected and properly interpreted, the form of the model, using x and x^2 to capture the curvilinearity, may be inappropriate. In particular, curvilinearity is often a consequence of a multiplicative relationship between y and x, in which case modelling the relationship by using logarithmic transformations is appropriate. For details, see *Multiplicative Regression Models* (893-013).