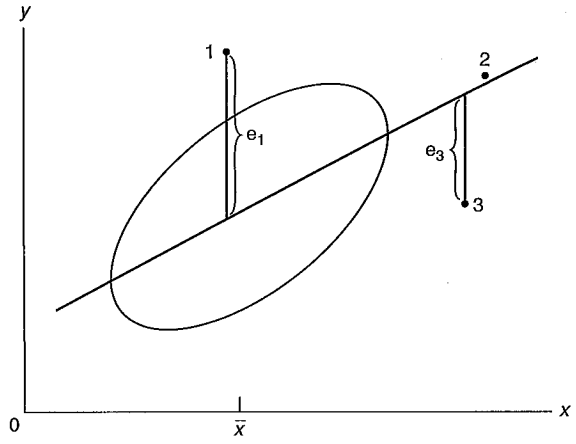


Figure 6.8: *Illustration of the Effect of Outliers*

Other distance measures have been proposed (Chatterjee and Hadi, 2006). The Welsch–Kuh distance measure (also called DFFITS) is the scaled distance between $\hat{Y}$ and $\hat{Y}(-i)$, i.e., $\hat{Y}$ derived with the ith observation deleted. Large values of DFFITS also indicate an influential observation. DFFITS tends to measure the influence on B and S^2 simultaneously.

A general lesson from the research work on outliers in regression analysis is that, when one examines either the scatter diagram of Y versus X or the plot of the residuals versus X, more attention should be given to the points that are outliers in both X and Y than to those that are only outliers in Y.

Lack of independence among residuals

When the observations can be ordered in time or place, plots of the residuals similar to those given in Figure 4.3 can be made and the discussion in Section 4.4 applies here directly. If the observations are independent, then successive residuals should not be appreciably correlated. The **serial** correlation, which is simply the correlation between successive residuals, can be used to assess lack of independence. For a sufficiently large N, the significance levels for the usual test $\rho = 0$ apply approximately to the serial correlation. Another test statistic available in some packaged programs is the Durbin–Watson statistic. The Durbin–Watson statistic is approximately equal to 2(1 – serial correlation). Thus when the serial correlation is zero, the Durbin–Watson statistic is close to two. The Durbin–Watson statistic is used to test whether the serial correlation is zero when it is assumed that the correlation between successive residuals

is restricted to a correlation between immediately adjacent residuals. Methods for dealing with correlated observations will be discussed in Chapter 18.

Normality of residuals

Some regression programs provide normal probability or quantile plots of the residuals to enable the user to decide whether the data approximate a normal distribution. If the residuals are not normally distributed, then the distribution of Y at each value of X is not normal.

For simple linear regression with the variable-X model, many researchers assess bivariate normality by examining the scatter diagram of Y versus X to see if the points approximately fall within an ellipse.

6.9 Robustness and transformations for regression

In this section we define the concept of robustness in statistical analysis, and we discuss the role of transformations in regression and correlation.

Robustness and assumptions

Regression and correlation analysis make certain assumptions about the population from which the data were obtained. A **robust analysis** is one that is useful even though all the assumptions are not met. For the purpose of fitting a straight line, we assume that the Y values are normally distributed, the population regression equation is linear in the range of concern, and the variance of Y is the same for all values of X. Linearity can be checked graphically, and transformations can help straighten out a nonlinear regression line.

The **assumption of homogeneity of variance** is not crucial for the resulting least squares line. In fact, the least squares estimates of α and β are unbiased whether or not the assumption is valid. However, if glaring irregularities of variance exist, weighted least squares can improve the fit. In this case the weights are chosen to be proportional to the inverse of the variance. For example, if the variance is a linear function of X, then the weight is $1/X$.

The **assumption of normality** of the Y values of each value of X is made only when tests of hypotheses are performed or confidence intervals are calculated. It is generally agreed in the statistical literature that slight departures from this assumption do not appreciably alter our inferences if the sample size is sufficiently large.

The **lack of randomness** in the sample can seriously invalidate our inferences. Confidence intervals are often optimistically narrow because the sample is not truly a random one from the whole population to which we wish to generalize.

In all of the preceding analyses **linearity** of the relationship between X

and Y was assumed. Thus careful examination of the scatter diagram should be the first step in any regression analysis. It is advisable to explore various transformations of Y and/or X if nonlinearity of the original measurements is apparent.

Transformations

The subject of **transformations** has been discussed in detail in the literature (e.g., Draper and Smith, 1998). In this subsection we present some typical graphs of the relationship between Y and X and some practical transformations.

In Chapter 4 we discussed the effects of transformations on the frequency distribution. There it was shown that taking the logarithm or square root of a number condensed the magnitude of larger numbers and stretched the magnitude of values less than one. Conversely, raising a number to a power greater than one stretches the large values and condenses the values less than one. These properties are useful in selecting the appropriate transformation to straighten out a nonlinear graph of one variable as a function of another.

Typical regression curves that are not linear in X can be viewed as one of the quadrants of Figure 6.9a (see the classic reference by Mosteller and Tukey, 1977). A very common case is illustrated in Figure 6.9b, which is represented by the fourth quadrant of the circle in Figure 6.9a. For example, the curve in Figure 6.9b might be made linear by transforming X to $\log X$, to $-1/X$, or to $X^{1/2}$. Another possibility would be to transform Y to Y^2. The other three cases are also indicated in Figure 6.9a. The remaining quadrants are interpreted in a similar fashion.

Other transformations could also be attempted, such as powers other than those indicated. It may also be useful to first add or subtract a constant from all values of X or Y and then take a power or logarithms. For example, sometimes taking $\log X$ does not straighten out the curve sufficiently. Subtracting a constant C (which must be smaller than the smallest X value) and then taking the logarithm has a greater effect.

The availability of packaged programs greatly facilitates the choice of an appropriate transformation. New variables can be created that are functions of the original variables, and scatter diagrams can be obtained of the new transformed variables. Visual inspection will often indicate the best transformation. Also, the magnitude of the correlation coefficient r will indicate the best linear fit since it is a measure of linear association. Attention should be paid to transformations that are commonly used in the field of application and that have a particular scientific basis or physical rationale.

Once the transformation is selected, all subsequent estimates and tests are performed in terms of the transformed values. Since the variable to be predicted is usually the dependent variable Y, transforming Y can complicate the interpretation of the resulting regression equation more than if X is transformed. For example, if $\log X$ is used instead of X, the resulting equation is

$$Y = A + B \log_{10} X$$

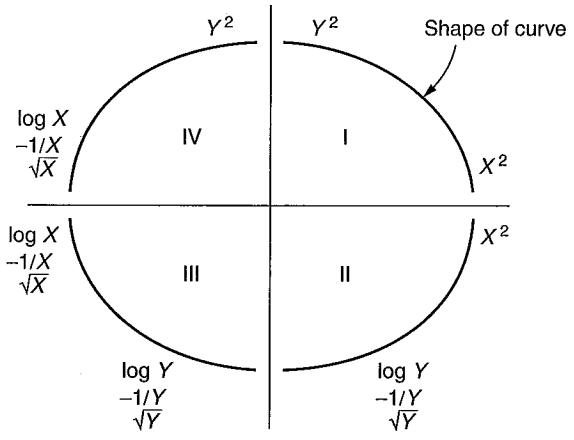
This equation presents no problems in interpreting the predicted values of Y, and most investigators accept the transformation of $\log_{10} X$ as reasonable in certain situations.

However, if $\log Y$ is used instead of Y, the resulting equation is

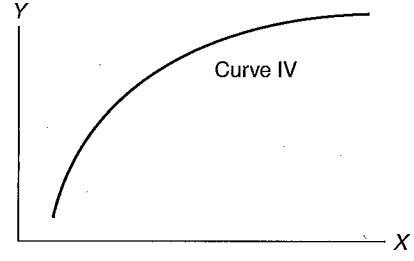
$$\log_{10} Y = A + BX$$

Then the predicted value of Y, say Y^*, must be detransformed, that is,

$$Y^* = 10^{A-BX}$$



a. Curves Not Linear in X



b. Detail of Fourth Quadrant

Figure 6.9: *Choice of Transformation: Typical Curves and Appropriate Transformation*

Thus slight biases in fitting log Y could be detransformed into large biases in predicting Y. For this reason most investigators look for transformations of X first.

6.10 Other types of regression

In this section we discuss three options available from computer programs: regression through the origin, weighted regression, and loess curves.

Regression through the origin

Sometimes an investigator is convinced that the **regression line** should pass **through the origin**. In this case the appropriate model is simply the mean of

$$Y = \beta X$$

That is, the intercept is forced to be zero. The programs usually give the option of using this model and estimate β as

$$B = \frac{\sum XY}{\sum X^2}$$

To test $H_0 : \beta = \beta_0$, the test statistic is

$$t = \frac{B - \beta_0}{S/(\sum X^2)^{1/2}}$$

where

$$S = \left[\frac{\sum (Y - BX)^2}{N-1}\right]^{1/2}$$

and t has $N-1$ degrees of freedom.

Weighted least squares regression

The investigator may also request a **weighted least squares regression line**. In weighted least squares each observation is given an individual weight reflecting its importance or degree of variability. There are three common situations in which a weighted regression line is appropriate.

1. The variance of the distribution at a given X is a function of the X value. An example of this situation was shown in Figure 6.7b.
2. Each Y observation is, in fact, the mean of several determinations, and that number varies from one value of X to another.

3. The investigator wishes to assign different levels of importance to different points. For example, data from different countries could be weighted either by the size of the population or by the perceived accuracy of the data.

In case 1 the weights are the inverse of the variances of the point. In case 2 the weights are the number of determinations at each X point. In case 3 the investigator must make up numerical weights to reflect the perception of importance or accuracy of the data points.

In weighted least squares regression the estimates of α and β and other statistics are adjusted to reflect these special characteristics of the observations. In most situations the weights will not affect the results appreciably unless they are quite different from each other. Since it is considerably more work to compute a weighted least squares regression equation, it is recommended that one of the computer programs listed in Section 6.12 be used, rather than hand calculations.

Loess curves

One use of weighted regression is to fit what is called **loess** curves. Loess is an abbreviation for local regression. Alternatively, the word **lowess** is also used. Lowess stands for locally weighted regression scatter plot smoothing. Loess curves are especially useful in illustrating the relationship between X and Y when it obviously is not a straight line, and no transformation can be found that results in a straight line or a polynomial (see Section 7.8).

Computation of loess curves requires several steps. First, the X values are ordered from smallest to largest so that X_1 is the smallest and X_N is the largest. Then, using a subset of the X (and the corresponding Y) surrounding each X_i, the programs compute a predicted $\hat{Y}_i$. We have to decide how many points surrounding each X_i to use to predict each $\hat{Y}_i$. In Stata, in the lowess command, the percent of the total number of points to be used is called the bwidth (bandwidth). Centered subsets of $N \times$ bwidth observations are used; thus bwidth $\times$ 100 represents the percent of the total observations to be used. In S-PLUS, in the loess smoother command, the fraction of the corresponding parameter is called the span value. In SAS PROC LOESS the appropriate options are bucket and smooth. Both, SPSS and STATISTICA have a feature named lowess for this purpose. Note that Statistica does not provide an option or parameter to choose the bandwidth of values to be used.

Often it takes several attempts at an appropriate number of points to get the desired curve. If the curve is too jagged, we can smooth it out by taking more points. Usually a span value greater than one-fourth is recommended (see Cleveland, 1993).

For example, if we choose a bandwidth that requires nine points, the statistical program will compute a weighted regression line using the nearest nine points for each X_i in the sample. For example, for X_5 the nearest nine points