CHAPTER 4

REGRESSION DIAGNOSTICS: DETECTION OF MODEL VIOLATIONS

4.1 INTRODUCTION

We have stated the basic results that are used for making inferences about simple and multiple linear regression models in Chapters 2 and 3. The results are based on summary statistics that are computed from the data. In fitting a model to a given body of data, we would like to ensure that the fit is not overly determined by one or few observations. The distribution theory, confidence intervals, and tests of hypotheses outlined in Chapters 2 and 3 are valid and have meaning only if the standard regression assumptions are satisfied. These assumptions are stated in this chapter (Section 4.2). When these assumptions are violated the standard results quoted previously do not hold and an application of them may lead to serious error. We reemphasize that the prime focus of this book is on the detection and correction of violations of the basic linear model assumptions as a means of achieving a thorough and informative analysis of the data. This chapter presents methods for checking these assumptions. We will rely mainly on graphical methods as opposed to applying rigid numerical rules to check for model violations.

Regression Analysis by Example, Fourth Edition. By Samprit Chatterjee and Ali S. Hadi

4.2 THE STANDARD REGRESSION ASSUMPTIONS

In the previous two chapters we have given the least squares estimates of the regression parameters and stated their properties. The properties of least squares estimators and the statistical analysis presented in Chapters 2 and 3 are based on the following assumptions:

1. **Assumptions about the form of the model**: The model that relates the response Y to the predictors $X_1, X_2, \ldots, X_p$ is assumed to be linear in the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$, namely,

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \tag{4.1}$$

which implies that the ith observation can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \; i = 1, 2, \ldots, n. \tag{4.2}$$

We refer to this as the *linearity* assumption. Checking the linearity assumption in simple regression is easy because the validity of this assumption can be determined by examining the scatter plot of Y versus X. A linear scatter plot ensures linearity. Checking the linearity in multiple regression is more difficult due to the high dimensionality of the data. Some graphs that can be used for checking the linearity assumption in multiple regression are given later in this chapter. When the linearity assumption does not hold, transformation of the data can sometimes lead to linearity. Data transformation is discussed in Chapter 6.

2. **Assumptions about the errors**: The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ in (4.2) are assumed to be *independently and identically distributed* (iid) normal random variables each with mean zero and a common variance σ^2. Note that this implies four assumptions:

   - The error $\varepsilon_i, i = 1, 2, \ldots, n$, has a normal distribution. We refer to this as the *normality assumption*. The normality assumption is not as easily validated especially when the values of the predictor variables are not replicated. The validity of the normality assumption can be assessed by examination of appropriate graphs of the residuals, as we describe later in this chapter.
   - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have mean zero.
   - The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ have the same (but unknown) variance σ^2. This is the *constant variance assumption*. It is also known by other names such as the *homogeneity* or the *homoscedasticity* assumption. When this assumption does not hold, the problem is called the *heterogeneity* or the *heteroscedasticity* problem. This problem is considered in Chapter 7.

- The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent of each other (their pairwise covariances are zero). We refer to this as the *independent-errors assumption.* When this assumption does not hold, we have the *autocorrelation* problem. This problem is considered in Chapter 8.

3. **Assumptions about the predictors**: There are three assumptions concerning the predictor variables:

   - The predictor variables $X_1, X_2, \ldots, X_p$ are nonrandom, that is, the values $x_{1j}, x_{2j}, \ldots, x_{nj}; j = 1, 2, \ldots, p$, are assumed fixed or selected in advance. This assumption is satisfied only when the experimenter can set the values of the predictor variables at predetermined levels. It is clear that under nonexperimental or observational situations this assumption will not be satisfied. The theoretical results that are presented in Chapters 2 and 3 will continue to hold, but their interpretation has to be modified. When the predictors are random variables, all inferences are conditional, conditioned on the observed data. It should be noted that this conditional aspect of the inference is consistent with the approach to data analysis presented in this book. Our main objective is to extract the maximum amount of information from the available data.

   - The values $x_{1j}, x_{2j}, \ldots, x_{nj}; j = 1, 2, \ldots, p$, are measured without error. This assumption is hardly ever satisfied. The errors in measurement will affect the residual variance, the multiple correlation coefficient, and the individual estimates of the regression coefficients. The exact magnitude of the effects will depend on several factors, the most important of which are the standard deviation of the errors of measurement and the correlation structure among the errors. The effect of the measurement errors will be to increase the residual variance and reduce the magnitude of the observed multiple correlation coefficient. The effects of measurement errors on individual regression coefficients are more difficult to assess. The estimate of the regression coefficient for a variable is affected not only by its own measurement errors, but also by the measurement errors of other variables included in the equation.

     Correction for measurement errors on the estimated regression coefficients, even in the simplest case where all the measurement errors are uncorrelated, requires a knowledge of the ratio between the variances of the measurement errors for the variables and the variance of the random error. Since these quantities are seldom, if ever, known (particularly in the social sciences, where this problem is most acute), we can never hope to remove completely the effect of measurement errors from the estimated regression coefficients. If the measurement errors are not large compared to the random errors, the effect of measurement errors is slight. In interpreting the coefficients in such an analysis, this point should be remembered. Although there is some problem in the

estimation of the regression coefficients when the variables are in error, the regression equation may still be used for prediction. However, the presence of errors in the predictors decreases the accuracy of predictions. For a more extensive discussion of this problem, the reader is referred to Fuller (1987), Chatterjee and Hadi (1988), and Chi-Lu and Van Ness (1999).

- The predictor variables $X_1, X_2, \ldots, X_p$ are assumed to be linearly independent of each other. This assumption is needed to guarantee the uniqueness of the least squares solution (the solution of the normal equations in (A.2) in the Appendix to Chapter 3). If this assumption is violated, the problem is referred to as the *collinearity* problem This problem is considered in Chapters 9 and 10.

The first two of the above assumptions about the predictors cannot be validated, so they do not play a major role in the analysis. However, they do influence the interpretation of the regression results.

4. **Assumptions about the observations**: All observations are equally reliable and have approximately equal role in determining the regression results and in influencing conclusions.

A feature of the method of least squares is that small or minor violations of the underlying assumptions do not invalidate the inferences or conclusions drawn from the analysis in a major way. Gross violations of the model assumptions can, however, seriously distort conclusions. Consequently, it is important to investigate the structure of the residuals and the data pattern through graphs.

4.3 VARIOUS TYPES OF RESIDUALS

A simple and effective method for detecting model deficiencies in regression analysis is the examination of residual plots. Residual plots will point to serious violations in one or more of the standard assumptions when they exist. Of more importance, the analysis of residuals may lead to suggestions of structure or point to information in the data that might be missed or overlooked if the analysis is based only on summary statistics. These suggestions or cues can lead to a better understanding and possibly a better model of the process under study. A careful graphical analysis of residuals may often prove to be the most important part of the regression analysis.

As we have seen in Chapters 2 and 3, when fitting the linear model in (4.1) to a set of data by least squares, we obtain the fitted values,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}, \; i = 1, 2, \ldots, n, \tag{4.3}$$

and the corresponding *ordinary* least squares residuals,

$$e_i = y_i - \hat{y}_i, \; i = 1, 2, \ldots, n. \tag{4.4}$$

The fitted values in (4.3) can also be written in an alternative form as

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \ldots + p_{in}y_n, \quad i = 1, 2, \ldots, n, \tag{4.5}$$

where the p_{ij}'s are quantities that depend only on the values of the predictor variables (they do not involve the response variable). Equation (4.5) shows directly the relationship between the observed and predicted values. In simple regression, p_{ij} is given by

$$p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum(x_i - \bar{x})^2}. \tag{4.6}$$

In multiple regression the p_{ij}'s are elements of a matrix known as the *hat* or *projection* matrix, which is defined in (A.5) in the Appendix to Chapter 3.

When $i = j$, p_{ii} is the ith diagonal element of the projection matrix $\mathbf{P}$. In simple regression,

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}. \tag{4.7}$$

The value p_{ii} is called the *leverage* value for the ith observation because, as can be seen from (4.5), $\hat{y}_i$ is a weighted sum of all observations in Y and p_{ii} is the weight (leverage) given to y_i in determining the ith fitted value $\hat{y}_i$ (Hoaglin and Welsch, 1978). Thus, we have n leverage values and they are denoted by

$$p_{11}, p_{22}, \ldots, p_{nn}. \tag{4.8}$$

The leverage values play an important role in regression analysis and we shall often encounter them.

When the assumptions stated in Section 4.2 hold, the ordinary residuals, e_1, e_2, ..., e_n, defined in (4.4), will sum to zero, but they will not have the same variance because

$$Var(e_i) = \sigma^2(1 - p_{ii}), \tag{4.9}$$

where p_{ii} is the ith leverage value in (4.8), which depends on x_{i1}, x_{i2}, ..., x_{ip}. To overcome the problem of unequal variances, we standardize the ith residual e_i by dividing it by its standard deviation and obtain

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}. \tag{4.10}$$

This is called the ith *standardized residual* because it has mean zero and standard deviation 1. The standardized residuals depend on σ, the unknown standard deviation of ε. An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum(y_i - \hat{y}_i)^2}{n - p - 1} = \frac{\text{SSE}}{n - p - 1}, \tag{4.11}$$

where SSE is the sum of squares of the residuals. The number $n - p - 1$ in the denominator of (4.11) is called the *degrees of freedom* ($d.f.$). It is equal to the

number of observations, n, minus the number of estimated regression coefficients, $p+1$.

An alternative unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2_{(i)} = \frac{\text{SSE}_{(i)}}{(n-1)-p-1} = \frac{\text{SSE}_{(i)}}{n-p-2}, \tag{4.12}$$

where $\text{SSE}_{(i)}$ is the sum of squared residuals when we fit the model to the $(n-1)$ observations obtained by omitting the ith observation. Both $\hat{\sigma}^2$ and $\hat{\sigma}^2_{(i)}$ are unbiased estimates of σ^2.

Using $\hat{\sigma}$ as an estimate of σ in (4.10), we obtain

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}, \tag{4.13}$$

whereas using $\hat{\sigma}_{(i)}$ as an estimate of σ, we obtain

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-p_{ii}}}. \tag{4.14}$$

The form of residual in (4.13) is called the *internally studentized residual*, and the residual in (4.14) is called the *externally studentized residual*, because e_i is not involved in (external to) $\hat{\sigma}_{(i)}$. For simplicity of terminology and presentation, however, we shall refer to the studentized residuals as the standardized residuals.

The standardized residuals do not sum to zero, but they all have the same variance. The externally standardized residuals follow a t-distribution with $n-p-2$ degrees of freedom, but the internally standardized residuals do not. However, with a moderately large sample, these residuals should approximately have a standard normal distribution. The residuals are not strictly independently distributed, but with a large number of observations, the lack of independence may be ignored.

The two forms of residuals are related by

$$r_i^* = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}}, \tag{4.15}$$

hence one is a monotone transformation of the other. Therefore, for the purpose of residuals plots, it makes little difference as to which of the two forms of the standardized residuals is used. From here on, we shall use the internally standardized residuals in the graphs. We need not make any distinction between the internally and externally standardized residuals in our residual plots. Several graphs of the residuals are used for checking the regression assumptions.

4.4 GRAPHICAL METHODS

Graphical methods play an important role in data analysis. It is of particular importance in fitting linear models to data. As Chambers et al. (1983, p. 1)

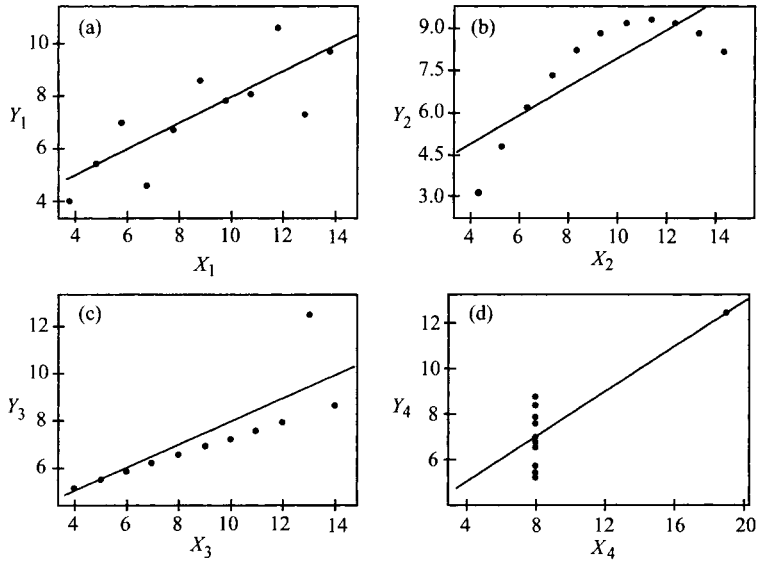

Figure 4.1 Plot of the data (X, Y) with the least squares fitted line for the Anscombe's quartet.

put it, "There is no single statistical tool that is as powerful as a well-chosen graph." Graphical methods can be regarded as exploratory tools. They are also an integral part of confirmatory analysis or statistical inference. Huber (1991) says, "Eye-balling can give diagnostic insights no formal diagnostics will ever provide." One of the best examples that illustrates this is the Anscombe's quartet, the four data sets given in Chapter 2 (Table 2.4). The four data sets are constructed by Anscombe (1973) in such a way that all pairs (Y, X) have identical values of descriptive statistics (same correlation coefficients, same regression lines, same standard errors, etc.), yet their pairwise scatter plots (reproduced in Figure 4.1 for convenience) give completely different scatters.

The scatter plot in Figure 4.1(a) indicates that a linear model may be reasonable, whereas the one in Figure 4.1(b) suggests a (possibly linearizable) nonlinear model. Figure 4.1(c) shows that the data follow a linear model closely except for one point which is clearly off the line. This point may be an outlier, hence it should be examined before conclusions can be drawn from the data. Figure 4.1(d) indicates either a deficient experimental design or a bad sample. For the point at $X = 19$, the reader can verify that (a) the residual at this point is always zero (with a variance of zero) no matter how large or small its corresponding value of Y and (b) if the point is removed, the least squares estimates based on the remaining points are no longer unique (except the vertical line, any line that passes through the average of the remaining points is a least squares line!). Observations which unduly influence regression results are called *influential observations*. The point at $X = 19$ is

therefore extremely influential because it alone determines both the intercept and the slope of the fitted line.

We have used the scatter plot here as an exploratory tool, but one can also use graphical methods to complement numerical methods in a confirmatory analysis. Suppose we wish to test whether there is a positive correlation between Y and X or, equivalently, if Y and X can be fitted by a positively sloped regression line. The reader can verify that the correlation coefficients are the same in all four data sets ($\text{Cor}(Y, X) = 0.80$) and all four data sets also have the same regression line ($Y = 3 + 0.5\,X$) with the same standard errors of the coefficients. Thus, based on these numerical summaries, one would reach the erroneous conclusion that all four data sets can be described by the same model. The underlying assumption here is that the relationship between Y and X is linear and this assumption does not hold here, for example, for the data set in Figure 4.1(b). Hence the test is invalid. The test for linear relationship, like other statistical methods, is based on certain underlying assumptions. Thus conclusions based on these methods are valid only when the underlying assumptions hold. It is clear from the above example that if analyses were solely based on numerical results, wrong conclusions will be reached.

Graphical methods can be useful in many ways. They can be used to:

1. Detect errors in the data (e.g., an outlying point may be a result of a typographical error),
2. Recognize patterns in the data (e.g., clusters, outliers, gaps, etc.),
3. Explore relationships among variables,
4. Discover new phenomena,
5. Confirm or negate assumptions,
6. Assess the adequacy of a fitted model,
7. Suggest remedial actions (e.g., transform the data, redesign the experiment, collect more data, etc.), and
8. Enhance numerical analyses in general.

This chapter presents some graphical displays useful in regression analysis. The graphical displays we discuss here can be classified into two (not mutually exclusive) classes:

- Graphs before fitting a model. These are useful, for example, in correcting errors in data and in selecting a model.
- Graphs after fitting a model. These are particularly useful for checking the assumptions and for assessing the goodness of the fit.

Our presentation draws heavily from Hadi (1993) and Hadi and Son (1997). Before examining a specific graph, consider what the graph should look like when the assumptions hold. Then examine the graph to see whether it is consistent with expectations. This will then confirm or disprove the assumption.

4.5 GRAPHS BEFORE FITTING A MODEL

The form of a model that represents the relationship between the response and predictor variables should be based on the theoretical background or the hypothesis to be tested. But if no prior information about the form of the model is available, the data may be used to suggest the model. The data should be examined thoroughly before a model is fitted. The graphs that one examines before fitting a model to the data serve as exploratory tools. Four possible groups of graphs are:

1. One-dimensional graphs,
2. Two-dimensional graphs,
3. Rotating plots, and
4. Dynamic graphs.

4.5.1 One-Dimensional Graphs

Data analysis usually begins with the examination of each variable in the study. The purpose is to have a general idea about the distribution of each individual variable. One of the following graphs may be used for examining a variable:

- Histogram
- Stem-and-leaf display
- Dot Plot
- Box Plot

The one-dimensional graphs serve two major functions. They indicate the distribution of a particular variable, whether the variable is symmetric or skewed. When a variable is very skewed it should be transformed. For a highly skewed variable a logarithmic transformation is recommended. Univariate graphs provide guidance on the question as to whether one should work with the original or with the transformed variables.

Univariate graphs also point out the presence of outliers in the variables. Outliers should be checked to see if they are due to transcription errors. No observation should be deleted at this stage. They should be noted as they may show up as troublesome points later.

4.5.2 Two-Dimensional Graphs

Ideally, when we have multidimensional data, we should examine a graph of the same dimension as that of the data. Obviously, this is feasible only when the number of variables is small. However, we can take the variables in pairs and look at the scatter plots of each variable versus each other variable in the data set. The

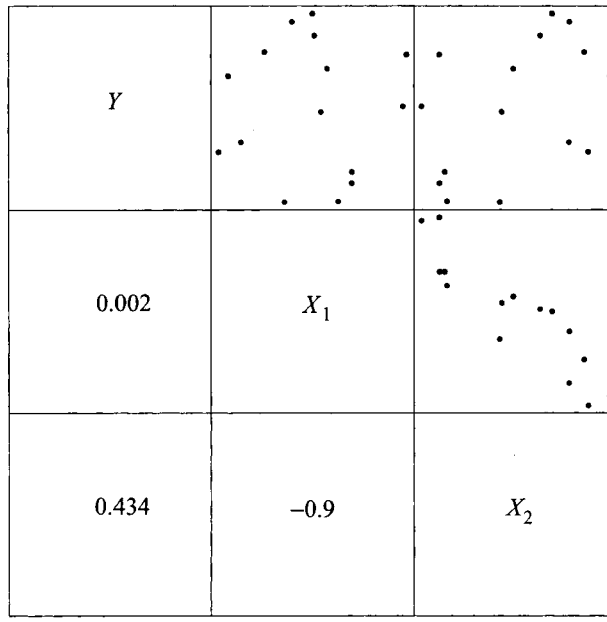

Figure 4.2 The plot matrix for Hamilton's data with the pairwise correlation coefficients.

purposes of these pairwise scatter plots are to explore the relationships between each pair of variables and to identify general patterns.

When the number of variables is small, it may be possible to arrange these pairwise scatter plots in a matrix format, sometimes referred to as the *draftsman*'s plot or the *plot* matrix. Figure 4.2 is an example of a plot matrix for one response and two predictor variables. The pairwise scatter plots are given in the upper triangular part of the plot matrix. We can also arrange the corresponding correlation coefficients in a matrix. The corresponding correlation coefficients are given in the lower triangular part of the plot matrix. These arrangements facilitate the examination of the plots. The pairwise correlation coefficients should always be interpreted in conjunction with the corresponding scatter plots. The reason for this is two-fold: (a) the correlation coefficient measures only linear relationships, and (b) the correlation coefficient is non-robust, that is, its value can be substantially influenced by one or two observations in the data.

What do we expect each of the graphs in the plot matrix to look like? In simple regression, the plot of Y versus X is expected to show a linear pattern. In multiple regression, however, the scatter plots of Y versus each predictor variable may or may not show linear patterns. Where the presence of a linear pattern is reassuring, the absence of such a pattern does not imply that our linear model is incorrect. An example is given below.

Table 4.1 Hamilton's (1987) Data

Y	X_1	X_2	Y	X_1	X_2
12.37	2.23	9.66	12.86	3.04	7.71
12.66	2.57	8.94	10.84	3.26	5.11
12.00	3.87	4.40	11.20	3.39	5.05
11.93	3.10	6.64	11.56	2.35	8.51
11.06	3.39	4.91	10.83	2.76	6.59
13.03	2.83	8.52	12.63	3.90	4.90
13.13	3.02	8.04	12.46	3.16	6.96
11.44	2.14	9.05			

Example: Hamilton's Data

Hamilton (1987) generates sets of data in such a way that Y depends on the predictor variables collectively but not individually. One such data set is given in Table 4.1. It can be seen from the plot matrix of this data (Figure 4.2) that no linear relationships exist in the plot of Y versus X_1 ($R^2 = 0$) and Y versus X_2 ($R^2 = 0.19$). Yet, when Y is regressed on X_1 and X_2 simultaneously, we obtain an almost perfect fit. The reader can verify that the following fitted equations are obtained:

$$\hat{Y} = 11.989 + 0.004\,X_1; \qquad t\text{-test} = 0.009; \qquad R^2 = 0.0,$$
$$\hat{Y} = 10.632 + 0.195\,X_2; \qquad t\text{-test} = 1.74; \qquad R^2 = 0.188,$$
$$\hat{Y} = -4.515 + 3.097\,X_1 + 1.032\,X_2; \qquad F\text{-test} = 39222; \qquad R^2 = 1.0.$$

The first two equations indicate that Y is related to neither X_1 nor X_2 individually, yet X_1 and X_2 predict Y almost perfectly. Incidentally, the first equation produces a negative value for the adjusted R^2, $R_a^2 = -0.08$.

The scatter plots that should look linear in the plot matrix are the plots of Y versus each predictor variable after adjusting for all other predictor variables (that is, taking the linear effects of all other predictor variables out). Two types of these graphs known as the *added-variable plot* and the *residual plus component plot*, are presented in Section 4.12.1.

The pairwise scatter plot of the predictors should show no linear pattern (ideally, we should see no discernible pattern, linear or otherwise) because the predictors are assumed to be linearly independent. In Hamilton's data, this assumption does not hold because there is a clear linear pattern in the scatter plot of X_1 versus X_2 (Figure 4.2). We should caution here that the absence of linear relationships in these scatter plots does not imply that the entire set of predictors are linearly independent. The linear relationship may involve more than two predictor variables. Pairwise scatter plots will fail to detect such a multivariate relationship. This multicollinearity problem will be dealt with in Chapters 9 and 10.

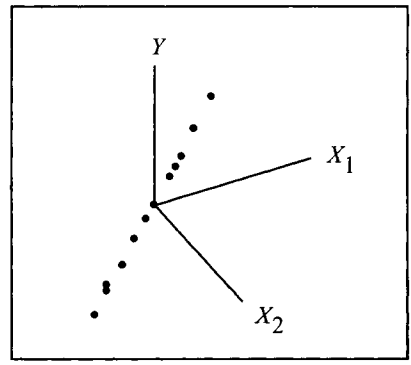

Figure 4.3 Rotating plot for Hamilton's data.

4.5.3 Rotating Plots

Recent advances in computer hardware and software have made it possible to plot data of three or more dimensions. The simplest of these plots is the three-dimensional rotating plot. The rotating plot is a scatter plot of three variables in which the points can be rotated in various directions so that the three-dimensional structure becomes apparent. Describing rotating plots in words does not do them justice. The real power of rotation can be felt only when one watches a rotating plot in motion on a computer screen. The motion can be stopped when one sees an interesting view of the data. For example, in the Hamilton's data we have seen that X_1 and X_2 predict Y almost perfectly. This finding is confirmed in the rotating plot of Y against X_1 and X_2. When this plot is rotated, the points fall on an almost perfect plane. The plot is rotated until an interesting direction is found. Figure 4.3 shows one such direction, where the plane is viewed from an angle that makes the scatter of points seem to fall on a straight line.

4.5.4 Dynamic Graphs

Dynamic graphics are an extraordinarily useful tool for exploring the structure and relationships in multivariate data. In a dynamic graphics environment the data analyst can go beyond just looking at a static graph. The graphs can be manipulated and the changes can be seen instantaneously on the computer screen. For example, one can make two or more three-dimensional rotating plots then use dynamic graphical techniques to explore the structure and relationships in more than three dimensions. Articles and books have been written about the subject, and many statistical software programs include dynamic graphical tools (e.g., rotating, brushing, linking, etc.). We refer the interested reader to Becker, Cleveland, and Wilks (1987), and Velleman (1999).

4.6 GRAPHS AFTER FITTING A MODEL

The graphs presented in the previous section are useful in data checking and the model formulation steps. The graphs after fitting a model to the data help in checking the assumptions and in assessing the adequacy of the fit of a given model. These graphs can be grouped into the following classes:

1. Graphs for checking the linearity and normality assumptions,
2. Graphs for the detection of outliers and influential observations, and
3. Diagnostic plots for the effect of variables.

4.7 CHECKING LINEARITY AND NORMALITY ASSUMPTIONS

When the number of variables is small, the assumption of linearity can be checked by interactively and dynamically manipulating the plots discussed in the previous section. The task of checking the linearity assumption becomes difficult when the number of variables is large. However, one can check the linearity and normality assumptions by examining the residuals after fitting a given model to the data.

The following plots of the standardized residuals can be used to check the linearity and normality assumptions:

1. *Normal probability plot of the standardized residuals:* This is a plot of the ordered standardized residuals versus the so-called *normal scores*. The normal scores are what we would expect to obtain if we take a sample of size n from a standard normal distribution. If the residuals are normally distributed, the ordered residuals should be approximately the same as the ordered normal scores. Under normality assumption, this plot should resemble a (nearly) straight line with an intercept of zero and a slope of one (these are the mean and the standard deviation of the standardized residuals, respectively).

2. *Scatter plots of the standardized residual against each of the predictor variables:* Under the standard assumptions, the standardized residuals are uncorrelated with each of the predictor variables. If the assumptions hold, this plot should be a random scatter of points. Any discernible pattern in this plot may indicate violation of some assumptions. If the linearity assumption does not hold, one may observe a plot like the one given in Figure 4.4(a). In this case a transformation of the Y and/or the particular predictor variable may be necessary to achieve linearity. A plot that looks like Figure 4.4(b), may indicate heterogeneity of variance. In this case a transformation of the data that stabilizes the variance may be needed. Several types of transformations for the corrections of some model deficiencies are described in Chapter 6.

3. *Scatter plot of the standardized residual versus the fitted values:* Under the standard assumptions, the standardized residuals are also uncorrelated with

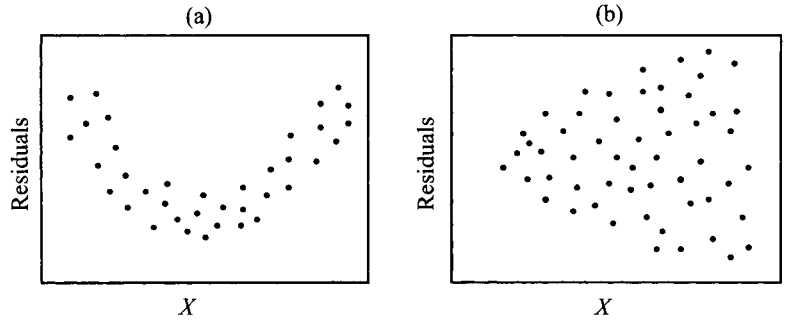

Figure 4.4 Two scatter plots of residuals versus X illustrating violations of model assumptions: (a) a pattern indicating nonlinearity; and (b) a pattern indicating heterogeneity.

the fitted values; therefore, this plot should also be a random scatter of points. In simple regression, the plots of standardized residuals against X and against the fitted values are identical.

4. *Index plot of the standardized residuals:* In this diagnostic plot we display the standardized residuals versus the observation number. If the order in which the observations were taken is immaterial, this plot is not needed. However, if the order is important (e.g., when the observations are taken over time or there is a spatial ordering), a plot of the residuals in serial order may be used to check the assumption of independence of the errors. Under the assumption of independent errors, the points should be scattered randomly within a horizontal band around zero.

4.8 LEVERAGE, INFLUENCE, AND OUTLIERS

In fitting a model to a given body of data, we would like to ensure that the fit is not overly determined by one or few observations. Recall, for example, that in the Anscombe's quartet data, the straight line for the data set in Figure 4.1(d) is determined entirely by one point. If the extreme point were to be removed, a very different line would result. When we have several variables, it is not possible to detect such a situation graphically. We would, however, like to know the existence of such points. It should be pointed out that looking at residuals in this case would be of no help, because the residual for this point is zero! The point is therefore not an outlier because it does not have a large residual, but it is a very influential point.

A point is an *influential* point if its deletion, singly or in combination with others (two or three), causes substantial changes in the fitted model (estimated coefficients, fitted values, t-tests, etc.). Deletion of any point will in general cause changes in the fit. We are interested in detecting those points whose deletion cause large changes (i.e., they exercise undue influence). This point is illustrated by an example.

Table 4.2 New York Rivers Data: The t-tests for the Individual Coefficients

	Observations Deleted		
Test	None	Neversink	Hackensack
t_0	1.40	1.21	2.08
t_1	0.39	0.92	0.25
t_2	−0.93	−0.74	−1.45
t_3	−0.21	−3.15	4.08
t_4	1.86	4.45	0.66

Example: New York Rivers Data

Consider the New York Rivers data described in Section 1.3.5 and given in Table 1.9. Let us fit a linear model relating the mean nitrogen concentration, Y, and the four predictor variables representing land use:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon. \tag{4.16}$$

Table 4.2 shows the regression coefficients and the t-tests for testing the significance of the coefficients for three subsets of the data. The second column in Table 4.2 gives the regression results based on all 20 observations (rivers). The third column gives the results after deleting the Neversink river (number 4). The fourth column gives the results after deleting the Hackensack river (number 5).

Note the striking difference among the regression outputs of three data sets that differ from each other by only one observation! Observe, for example, the values of the t-test for β_3. Based on all data, the test is insignificant, based on the data without the Neversink river, it is significantly negative, and based on the data without the Hackensack river, it is significantly positive. Only one observation can lead to substantially different results and conclusions! The Neversink and Hackensack rivers are called influential observations because they influence the regression results substantially more than other observations in the data. Examining the raw data in Table 1.9, one can easily identify the Hackensack river because it has an unusually large value for X_3 (percentage of residential land) relative to the other values for X_3. The reason for this large value is that the Hackensack river is the only urban river in the data due to its geographic proximity to New York City with its high population density. The other rivers are in rural areas. Although the Neversink river is influential (as can be seen from Table 4.2), it is not obvious from the raw data that it is different from the other rivers in the data.

It is therefore important to identify influential observations if they exist in data. We describe methods for the detection of influential observations. Influential observations are usually outliers in either the response variable Y or the predictor variable (the X-space).

4.8.1 Outliers in the Response Variable

Observations with large standardized residuals are outliers in the response variable because they lie far from the fitted equation in the Y-direction. Since the standardized residuals are approximately normally distributed with mean zero and a standard deviation 1, points with standardized residuals larger than 2 or 3 standard deviation away from the mean (zero) are called *outliers*. Outliers may indicate a model failure for those points. They can be identified using formal testing procedures (see, e.g., Hawkins (1980), Barnett and Lewis (1994), Hadi and Simonoff (1993), and Hadi and Velleman (1997) and the references therein) or through appropriately chosen graphs of the residuals, the approach we adopt here. The pattern of the residuals is more important than their numeric values. Graphs of residuals will often expose gross model violations when they are present. Studying residual plots is one of the main tools in our analysis.

4.8.2 Outliers in the Predictors

Outliers can also occur in the predictor variables (the X-space). They can also affect the regression results. The leverage values p_{ii}, described earlier, can be used to measure outlyingness in the X-space. This can be seen from an examination of the formula for p_{ii} in the simple regression case given in (4.7), which shows that the farther a point is from $\bar{x}$, the larger the corresponding value of p_{ii}. This is also true in multiple regression. Therefore, p_{ii} can be used as a measure of outlyingness in the X-space because observations with large values of p_{ii} are outliers in the X-space (i.e., compared to other points in the space of the predictors). Observations that are outliers in the X-space (e.g., the point with the largest value of X_4 in Figure 4.1(d)) are known as *high leverage* points to distinguish them from observations that are outliers in the response variable (those with large standardized residuals).

The leverage values possess several interesting properties (see Dodge and Hadi (1999) and Chatterjee and Hadi (1988), Chapter 2, for a comprehensive discussion). For example, they lie between zero and 1 and their average value is $(p + 1)/n$. Points with p_{ii} greater than $2(p + 1)/n$ (twice the average value) are generally regarded as points with high leverage (Hoaglin and Welsch, 1978).

In any analysis, points with high leverage should be flagged and then examined to see if they are also influential. A plot of the leverage values (e.g., index plot, dot plot, or a box plot) will reveal points with high leverage if they exist.

4.8.3 Masking and Swamping Problems

The standardized residuals provide valuable information for validating linearity and normality assumptions and for the identification of outliers. However, analyses that are based on residuals alone may fail to detect outliers and influential observations for the following reasons:

1. *The presence of high leverage points*: The ordinary residuals, e_i, and leverage values, p_{ii} are related by

$$p_{ii} + \frac{e_i^2}{\text{SSE}} \leq 1, \tag{4.17}$$

where SSE is the residual sum of squares. This inequality indicates that high leverage points (points with large values of p_{ii}) tend to have small residuals. For example, the point at $X = 19$ in Figure 4.1(d) is extremely influential even though its residual is identically zero. Therefore, in addition to an examination of the standardized residuals for outliers, an examination of the leverage values is also recommended for the identification of troublesome points.

2. *The masking and swamping problems*: Masking occurs when the data contain outliers but we fail to detect them. This can happen because some of the outliers may be hidden by other outliers in the data. Swamping occurs when we wrongly declare some of the non-outlying points as outliers. This can occur because outliers tend to pull the regression equation toward them, hence make other points lie far from the fitted equation. Thus, masking is a false negative decision whereas swamping is a false positive. An example of a data set in which masking and swamping problems are present is given below. Methods which are less susceptible to the masking and swamping problems than the standardized residuals and leverage values are given in Hadi and Simonoff (1993) and the references therein.

For the above reasons, additional measures of the influence of observations are needed. Before presenting these methods, we illustrate the above concepts using a real-life example.

Example: New York Rivers Data

Consider the New York rivers data, but now for illustrative purpose, let us consider fitting the simple regression model

$$Y = \beta_0 + \beta_4 X_4 + \varepsilon, \tag{4.18}$$

relating the mean nitrogen concentration, Y, to the percentage of land area in either industrial or commercial use, X_4. The scatter plot of Y versus X_4 together with the corresponding least squares fitted line are given in Figure 4.5. The corresponding standardized residuals, r_i, and the leverage values, p_{ii}, are given in Table 4.3 and their respective index plots are shown in Figure 4.6. In the index plot of the standardized residuals all the residuals are small indicating that there are no outliers in the data. This is a wrong conclusion because there are two clear outliers in the data as can be seen in the scatter plot in Figure 4.5. Thus masking has occurred! Because of the relationship between leverage and residual in (4.17), the Hackensack river with its large value of $p_{ii} = 0.67$, has a small residual. While a small value

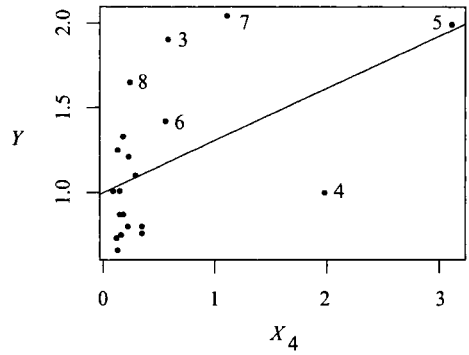

Figure 4.5 New York Rivers Data: Scatter plot of Y versus X_4.

Table 4.3 New York Rivers Data: The Standardized Residuals, r_i, and the Leverage Values, p_{ii}, From Fitting Model 4.18

Row	r_i	p_{ii}	Row	r_i	p_{ii}
1	0.03	0.05	11	0.75	0.06
2	−0.05	0.07	12	−0.81	0.06
3	1.95	0.05	13	−0.83	0.06
4	−1.85	0.25	14	−0.83	0.05
5	0.16	0.67	15	−0.94	0.05
6	0.67	0.05	16	−0.48	0.06
7	1.92	0.08	17	−0.72	0.06
8	1.57	0.06	18	−0.50	0.06
9	−0.10	0.06	19	−1.03	0.06
10	0.38	0.06	20	0.57	0.06

of the residual is desirable, the reason for the small value of the residual here is not due to a good fit; it is due to the fact that observation 5 is a high-leverage point and, in collaboration with observation 4, they pull the regression line toward them.

A commonly used cutoff value for p_{ii} is $2(p+1)/n = 0.2$ (Hoaglin and Welsch, 1978). Accordingly, two points (Hackensack, $p_{ii} = 0.67$, and Neversink, $p_{ii} = 0.25$) that we have seen previously stand out in the scatter plot of points in Figure 4.5, are flagged as high leverage points as can be seen in the index plot of p_{ii} in Figure 4.6(b), where the two points are far from the other points. This example shows clearly that looking solely at residual plots is inadequate.

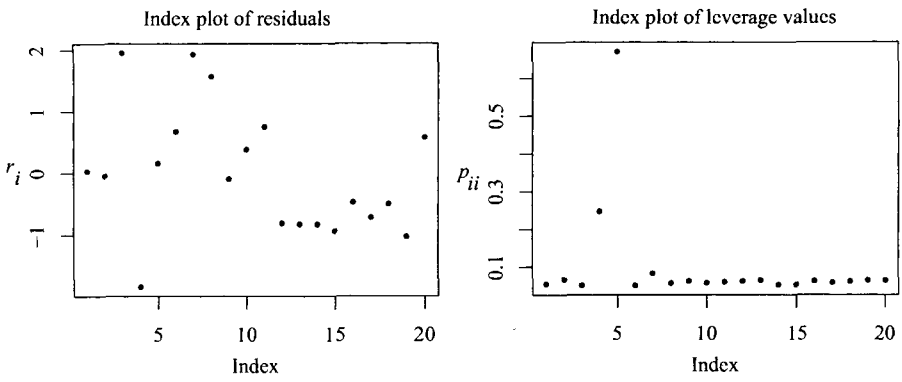

Figure 4.6 New York Rivers Data: Index plots of the standardized residuals, r_i, and the leverage values, p_{ii}.

4.9 MEASURES OF INFLUENCE

The influence of an observation is measured by the effects it produces on the fit when it is deleted in the fitting process. This deletion is almost always done one point at a time. Let $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}, \ldots, \hat{\beta}_{p(i)}$ denote the regression coefficients obtained when the ith observation is deleted ($i = 1, 2, \ldots, n$). Similarly, let $\hat{y}_{1(i)}, \hat{y}_{2(i)}, \ldots, \hat{y}_{n(i)}$, and $\hat{\sigma}^2_{(i)}$ be the predicted values and residual mean square when we drop the ith observation. Note that

$$\hat{y}_{m(i)} = \hat{\beta}_{0(i)} + \hat{\beta}_{1(i)} x_{m1} + \ldots + \hat{\beta}_{p(i)} x_{mp} \tag{4.19}$$

is the fitted value for observation m when the fitted equation is obtained with the ith observation deleted. Influence measures look at differences produced in quantities such as $(\hat{\beta}_j - \hat{\beta}_{j(i)})$ or $(\hat{y}_j - \hat{y}_{j(i)})$. There are numerous measures of influence in the literature, and the reader is referred to one of the books for details: Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985), and Chatterjee and Hadi (1988). Here we give three of these measures.

4.9.1 Cook's Distance

An influence measure proposed by Cook (1977) is widely used. *Cook's distance* measures the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the ith observation, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation. Accordingly, Cook's distance measures the influence of the ith observation by

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2 (p+1)}, \quad i = 1, 2, \ldots, n. \tag{4.20}$$

It can be shown that C_i can be expressed as

$$C_i = \frac{r_i^2}{p+1} \times \frac{p_{ii}}{1-p_{ii}}, \; i = 1, 2, \ldots, n. \tag{4.21}$$

Thus, Cook's distance is a multiplicative function of two basic quantities. The first is the square of the standardized residual, r_i, defined in (4.13) and the second is the so-called *potential* function $p_{ii}/(1-p_{ii})$, where p_{ii} is the leverage of the ith observation introduced previously. If a point is influential, its deletion causes large changes and the value of C_i will be large. Therefore, a large value of C_i indicates that the point is influential. It has been suggested that points with C_i values greater than the 50% point of the F distribution with $p+1$ and $(n-p-1)$ degrees of freedom be classified as influential points. A practical operational rule is to classify points with C_i values greater than 1 as being influential. Rather than using a rigid cutoff rule, we suggest that all C_i values be examined graphically. A dot plot or an index plot of C_i is a useful graphical device. When the C_i values are all about the same, no action need be taken. On the other hand, if there are data points with C_i values that stand out from the rest, these points should be flagged and examined. The model may then be refitted without the offending points to see the effect of these points.

4.9.2 Welsch and Kuh Measure

A measure similar to Cook's distance has been proposed by Welsch and Kuh (1977) and named DFITS. It is defined as

$$\text{DFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{p_{ii}}}, \; i = 1, 2, \ldots, n. \tag{4.22}$$

Thus, DFITS_i is the scaled difference between the ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation. The difference is scaled by $\hat{\sigma}_{(i)}\sqrt{p_{ii}}$. It can be shown that DFITS_i can be written as

$$\text{DFITS}_i = r_i^* \sqrt{\frac{p_{ii}}{1-p_{ii}}}, \; i = 1, 2, \ldots, n, \tag{4.23}$$

where r_i^* is the standardized residual defined in (4.14). DFITS_i corresponds to $\sqrt{C_i}$ when the normalization is done by using $\hat{\sigma}_{(i)}$ instead of $\hat{\sigma}$. Points with $|\text{DFITS}_i|$ larger than $2\sqrt{(p+1)/(n-p-1)}$ are usually classified as influential points. Again, instead of having a strict cutoff value, we use the measure to sort out points of abnormally high influence relative to other points on a graph such as the index plot, the dot plot, or the box plot. There is not much to choose between C_i and DFITS_i – both give similar answers because they are functions of the residual and leverage values. Most computer software will give one or both of the measures, and it is sufficient to look at only one of them.

4.9.3 Hadi's Influence Measure

Hadi (1992) proposed a measure of the influence of the ith observation based on the fact that influential observations are outliers in either the response variable or in the predictors, or both. Accordingly, the influence of the ith observation can be measured by

$$H_i = \frac{p_{ii}}{1-p_{ii}} + \frac{p+1}{1-p_{ii}} \frac{d_i^2}{1-d_i^2}, \; i = 1, 2, \ldots, n, \tag{4.24}$$

where $d_i = e_i/\sqrt{\text{SSE}}$ is the so-called *normalized residual*. The first term on the right-hand-side of (4.24) is the potential function which measures outlyingness in the X-space. The second term is a function of the residual, which measures outlyingness in the response variable. It can be seen that observations will have large values of H_i if they are outliers in the response and/or the predictor variables, that is, if they have large values of r_i, p_{ii}, or both. The measure H_i does not focus on a specific regression result, but it can be thought of as an overall general measure of influence which depicts observations that are influential on at least one regression result.

Note that C_i and DFITS_i are multiplicative functions of the residuals and leverage values, whereas H_i is an additive function. The influence measure H_i can best be examined graphically in the same way as Cook's distance and Welsch and Kuh measure.

Example: New York Rivers Data

Consider again fitting the simple regression model in (4.18), which relates the mean nitrogen concentration, Y, to the percentage of land area in commercial/industrial use, X_4. The scatter plot of Y versus X_4 and the corresponding least-squares regression are given in Figure 4.5. Observations 4 (the Neversink river) and 5 (the Hackensack river) are located far from the bulk of other data points in Figure 4.5. Also observations 7, 3, 8, and 6 are somewhat sparse in the upper-left region of the graph. The three influence measures discussed above which result from fitting model (4.18) are shown in Table 4.4, and the corresponding index plots are shown in Figure 4.7. No value of C_i exceeds it cutoff value of 1. However, the index plot of C_i in Figure 4.7(a) shows clearly that observation number 4 (Neversink) should be flagged as an influential observation. This observation also exceeds its DFITS_i cutoff value of $2\sqrt{(p+1)/(n-p-1)} = 2/3$. As can be seen from Figure 4.7, observation number 5 (Hackensack) was not flagged by C_i or by DFITS_i. This is due to the small value of the residual because of its high leverage and to the multiplicative nature of the measure. The index plot of H_i in Figure 4.7(c)

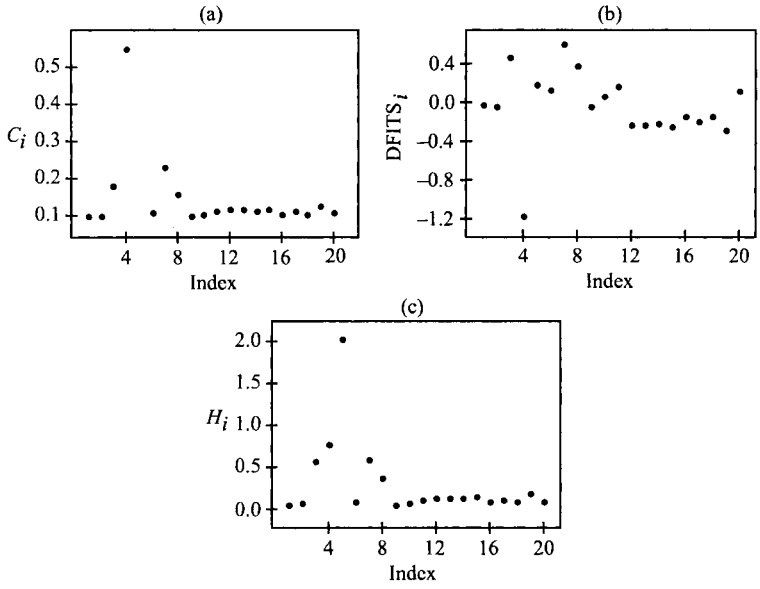

Figure 4.7 New York Rivers data: Index plots of influence measures (a) Cook's distance, C_i, (b) Welsch and Kuh measure, DFITS$_i$, and (c) Hadi's influence measure H_i.

Table 4.4 New York Rivers Data. Influence Measures From Fitting Model 4.18: Cook's Distance, C_i, Welsch and Kuh Measure, DFITS$_i$, and Hadi's Influence Measure H_i

Row	C_i	DFITS$_i$	H_i	Row	C_i	DFITS$_i$	H_i
1	0.00	0.01	0.06	11	0.02	0.19	0.13
2	0.00	−0.01	0.07	12	0.02	−0.21	0.14
3	0.10	0.49	0.58	13	0.02	−0.22	0.15
4	0.56	−1.14	0.77	14	0.02	−0.19	0.13
5	0.02	0.22	2.04	15	0.02	−0.22	0.16
6	0.01	0.15	0.10	16	0.01	−0.12	0.09
7	0.17	0.63	0.60	17	0.02	−0.18	0.12
8	0.07	0.40	0.37	18	0.01	−0.12	0.09
9	0.00	−0.02	0.07	19	0.04	−0.27	0.19
10	0.00	0.09	0.08	20	0.01	0.15	0.11

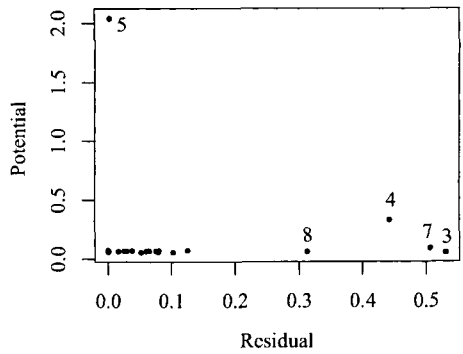

Figure 4.8 New York Rivers data: Potential-Residual plot.

indicates that observation number 5 (Hackensack) is the most influential one, followed by observation number 4 (Neversink), which is consistent with the scatter plot in Figure 4.5.

4.10 THE POTENTIAL-RESIDUAL PLOT

The formula for H_i in (4.24) suggests a simple graph to aid in classifying unusual observations as high-leverage points, outliers, or a combination of both. The graph is called the *potential-residual* (P-R) plot (Hadi, 1992) because it is the scatter plot of

$$\begin{array}{cccc} \text{Potential Function} & & & \text{Residual Function} \\ \dfrac{p_{ii}}{1-p_{ii}} & \text{versus} & \dfrac{p+1}{1-p_{ii}}\dfrac{d_i^2}{1-d_i^2} & . \end{array}$$

The P-R plot is related to the L-R (*leverage-residual*) plot suggested by Gray (1986) and McCulloch and Meeter (1983). The L-R plot is a scatter plot of p_{ii} versus d_i^2. For a comparison between the two plots, see Hadi (1992).

As an illustrative example, the P-R plot obtained from fitting model (4.18) is shown in Figure 4.8. Observation 5, which is a high-leverage point, is located by itself in the upper left corner of the plot. Four outlying observations (3, 7, 4, and 8) are located in the lower right area of the graph.

It is clear now that some individual data points may be flagged as outliers, leverage points, or influential points. The main usefulness of the leverage and influence measures is that they give the analyst a complete picture of the role played by different points in the entire fitting process. Any point falling in one of these categories should be carefully examined for accuracy (gross error, transcription error), relevancy (whether it belongs to the data set), and special significance

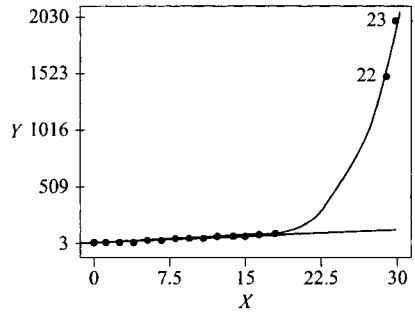

Figure 4.9 A scatter plot of population size, Y, versus time, X. The curve is obtained by fitting an exponential function to the full data. The straight line is the least squares line when observations 22 and 23 are are deleted.

(abnormal condition, unique situation). Outliers should always be scrutinized carefully. Points with high leverage that are not influential do not cause problems. High leverage points that are influential should be investigated because these points are outlying as far as the predictor variables are concerned and also influence the fit. To get an idea of the sensitivity of the analysis to these points, the model should be fitted without the offending points and the resulting coefficients examined.

4.11 WHAT TO DO WITH THE OUTLIERS?

Outliers and influential observations should not routinely be deleted or automatically down-weighted because they are not necessarily bad observations. On the contrary, if they are correct, they may be the most informative points in the data. For example, they may indicate that the data did not come from a normal population or that the model is not linear. To illustrate that outliers and influential observations can be the most informative points in the data, we use the exponential growth data described in the following example.

Example: Exponential Growth Data

Figure 4.9 is the scatter plot of two variables, the size of a certain population, Y, and time, X. As can be seen from the scatter of points, the majority of the points resemble a linear relationship between population size and time as indicated by the straight line in Figure 4.9. According to this model the two points 22 and 23 in the upper right corner are outliers. If these points, however, are correct, they are the only observations in the data set that indicate that the data follow a nonlinear (e.g., exponential) model, such as the one shown in the graph. Think of this as a population of bacteria which increases very slowly over a period of time. After a critical point in time, however, the population explodes.

What to do with outliers and influential observations once they are identified? Because outliers and influential observations can be the most informative observations in the data set, they should not be automatically discarded without justification. Instead, they should be examined to determine why they are outlying or influential. Based on this examination, appropriate corrective actions can then be taken. These corrective actions include: correction of error in the data, deletion or down-weighing outliers, transforming the data, considering a different model, and redesigning the experiment or the sample survey, collecting more data.

4.12 ROLE OF VARIABLES IN A REGRESSION EQUATION

As we have indicated, successive variables are introduced sequentially into a regression equation. A question that arises frequently in practice is: Given a regression model which currently contains p predictor variables, what are the effects of deleting (or adding) one of the variables from (or to) the model? Frequently, the answer is to compute the t-test for each variable in the model. If the t-test is large in absolute value, the variable is retained, otherwise the variable is omitted. This is valid only if the underlying assumptions hold. Therefore, the t-test should be interpreted in conjunction with appropriate graphs of the data. Two plots have been proposed that give this information visually and are often very illuminating. They can be used to complement the t-test in deciding whether one should retain or remove a variable in a regression equation. The first graph is called the *added-variable plot* and the second is the *residual plus component plot.*

4.12.1 Added-Variable Plot

The added-variable plot, introduced by Mosteller and Tukey (1977), enables us graphically to see the magnitude of the regression coefficient of the new variable that is being considered for inclusion. The slope of the least squares line representing the points in the plot is equal to the estimated regression coefficient of the new variable. The plot also shows data points which play key roles in determining this magnitude. We can construct an added-variable plot for each predictor variable X_j. The added-variable plot for X_j is essentially a graph of two different sets of residuals. The first is the residuals when Y is regressed on all predictor variables except X_j. We call this set the Y-residuals. The second set of residuals are obtained when we regress X_j (treated temporarily as a response variable) on all other predictor variables. We refer to this set as the X_j-residuals. Thus, the added-variable plot for X_j is simply a scatter plot of the

$$Y\text{-residuals versus } X_j\text{-residuals.}$$

Therefore, if we have p predictor variables available, we can construct p added-variable plots, one for each predictor.

Note that the Y-residuals in the added-variable plot for X_j represent the part of Y not explained by all predictors other than X_j. Similarly, the X_j-residuals

represent the part of X_j that is not explained by the other predictor variables. If a least squares regression line were fitted to the points in the added-variable plot for X_j, the slope of this line is equal to $\hat{\beta}_j$, the estimated regression coefficient of X_j when Y is regressed on all the predictor variables including X_j. This is an illuminating but equivalent interpretation of the partial regression coefficient as we have seen in Section 3.5.

The slope of the points in the plot gives the magnitude of the regression coefficient of the variable if it were brought into the equation. Thus, the stronger the linear relationship in the added-variable plot is, the more important the additional contribution of X_j to the regression equation already containing the other predictors. If the scatter of the points shows no marked slope, the variable is unlikely to be useful in the model. The scatter of the points will also indicate visually which of the data points are most influential in determining this slope and its corresponding t-test. The added-variable plot is also known as the *partial regression plot.*. We remark in passing that it is not actually necessary to carry out this fitting. These residuals can be obtained very simply from computations done in fitting Y on the full set of predictors. For a detailed discussion, see Velleman and Welsch (1981) and Chatterjee and Hadi (1988).

4.12.2 Residual Plus Component Plot

The residual plus component plot, introduced by Ezekiel (1924), is one of the earliest graphical procedures in regression analysis. It was revived by Larsen and McCleary (1972), who called it a *partial residual plot.* We are calling it a residual plus component plot, after Wood (1973), because this name is more self-explanatory.

The residual plus component plot for X_j is a scatter plot of

$$(e + \hat{\beta}_j X_j) \text{ versus } X_j,$$

where e is the ordinary least squares residuals when Y is regressed on all predictor variables and $\hat{\beta}_j$ is the coefficient of X_j in this regression. Note that $\hat{\beta}_j X_j$ is the contribution (component) of the jth predictor to the fitted values. As in the added-variable plot, the slope of the points in this plot is $\hat{\beta}_j$, the regression coefficient of X_j. Besides indicating the slope graphically, this plot indicates whether any nonlinearity is present in the relationship between Y and X_j. The plot can therefore suggest possible transformations for linearizing the data. The indication of nonlinearity is, however, not present in the added-variable plot because the horizontal scale in the plot is not the variable itself. Both plots are useful, but the residual plus component plot is more sensitive than the added-variable plot in detecting nonlinearities in the variable being considered for introduction in the model. The added-variable plot is, however, easier to interpret and points out the influential observations.

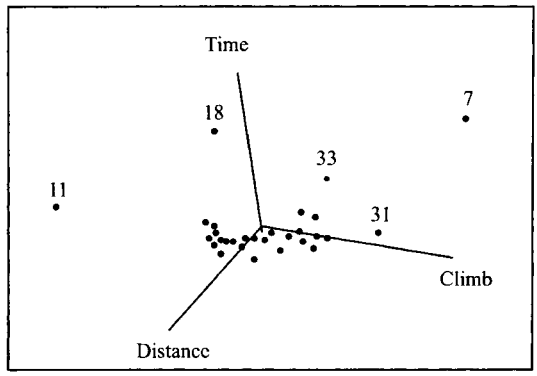

Figure 4.10 Rotating plot for the Scottish Hills Races data.

Example: The Scottish Hills Races Data

The Scottish Hills Races data consist of a response variable (record times, in seconds) and two explanatory variables (the distance in miles, and the climb in feet) for 35 races in Scotland in 1984. The data set is given in Table 4.5. Since this data set is three-dimensional, let us first examine a three-dimensional rotating plot of the data as an exploratory tool. An interesting direction in this rotating plot is shown in Figure 4.10. Five observations are marked in this plot. Clearly, observations 7 and 18 are outliers, they lie far away (in the direction of Time) from the plane suggested by the majority of other points. Observation 7 lies far away in the direction of Climb. Observations 33 and 31 are also outliers in the graph but to a lesser extent. While observations 11 and 31 are near the plane suggested by the majority of other points, they are located far from the rest of the points on the plane. (Observation 11 is far mainly in the direction of Distance and observation 31 is in the direction of Climb.) The rotating plot clearly shows that the data contain unusual points (outliers, high leverage points, and/or influential observations).

The fitted equation is

$$Time = -539.483 + 373.073\ Distance\ + 0.662888\ Climb. \qquad (4.25)$$

We wish to address the question: Does each of the predictor variables contribute significantly when the other variable is included in the model? The t-test for the two predictors are 10.3 and 5.39, respectively, indicating very high significance. This implies that the answer to the above question is in the affirmative for both variables. The validity of this conclusion can be enhanced by examining the corresponding added-variable and residual plus component plots. These are given in Figures 4.11 and 4.12, respectively. For example, in the added-variable plot for Distance in Figure 4.11(a), the quantities plotted on the ordinate axis are the residuals obtained from the regression of Time on Climb (the other predictor variable), and the quantities plotted on the abscissa are the residuals obtained from the regression

Table 4.5 Scottish Hills Races Data

Row	Race	Time	Distance	Climb
1	Greenmantle New Year Dash	965	2.5	650
2	Carnethy	2901	6	2500
3	Craig Dunain	2019	6	900
4	Ben Rha	2736	7.5	800
5	Ben Lomond	3736	8	3070
6	Goatfell	4393	8	2866
7	Bens of Jura	12277	16	7500
8	Cairnpapple	2182	6	800
9	Scolty	1785	5	800
10	Traprain Law	2385	6	650
11	Lairig Ghru	11560	28	2100
12	Dollar	2583	5	2000
13	Lomonds of Fife	3900	9.5	2200
14	Cairn Table	2648	6	500
15	Eildon Two	1616	4.5	1500
16	Cairngorm	4335	10	3000
17	Seven Hills of Edinburgh	5905	14	2200
18	Knock Hill	4719	3	350
19	Black Hill	1045	4.5	1000
20	Creag Beag	1954	5.5	600
21	Kildoon	957	3	300
22	Meall Ant-Suiche	1674	3.5	1500
23	Half Ben Nevis	2859	6	2200
24	Cow Hill	1076	2	900
25	North Berwick Law	1121	3	600
26	Creag Dubh	1573	4	2000
27	Burnswark	2066	6	800
28	Largo	1714	5	950
29	Criffel	3030	6.5	1750
30	Achmony	1257	5	500
31	Ben Nevis	5135	10	4400
32	Knockfarrel	1943	6	600
33	Two Breweries Fell	10215	18	5200
34	Cockleroi	1686	4.5	850
35	Moffat Chase	9590	20	5000

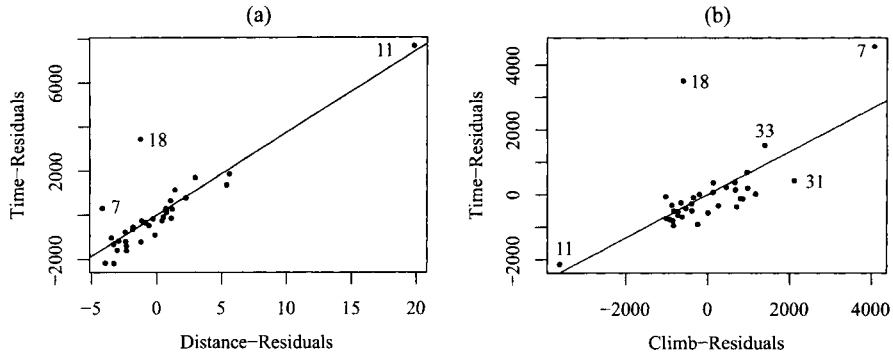

Figure 4.11 The Scottish Hills Races Data: Added-variable plots for (a) Distance and (b) Climb.

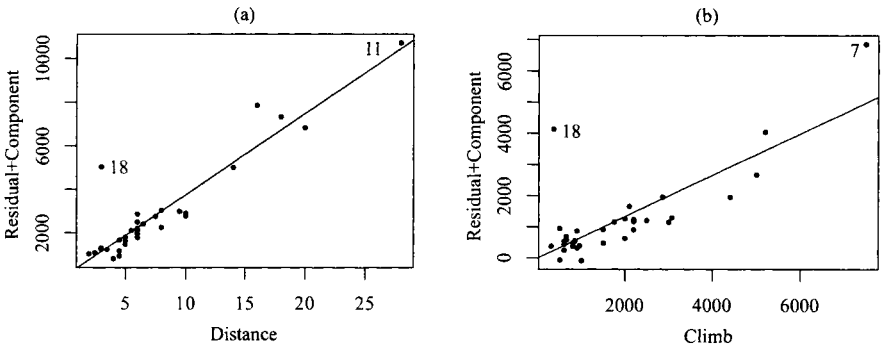

Figure 4.12 The Scottish Hills Races Data: Residual plus component plots for (a) Distance and (b) Climb.

of Distance on Climb. Similarly for the added-variable plot for Climb, the quantities plotted are the residuals obtained from the regression of Time on Distance and the residuals obtained from the regression of Climb on Distance.

It can be seen that there is strong linear trend in all four graphs supporting the conclusions reached by the above t-tests. The graphs, however, indicate the presence of some points that may influence our results and conclusions. Races 7, 11, and 18 clearly stand out. These points are marked on the graphs by their numbers. Races 31 and 33 are also suspects but to a lesser extent. An examination of the P-R plot obtained from the above fitted equation (Figure 4.13) classifies Race 11 as a high leverage point, Race 18 as an outlier, and Race 7 as a combination of both. These points should be scrutinized carefully before continuing with further analysis.

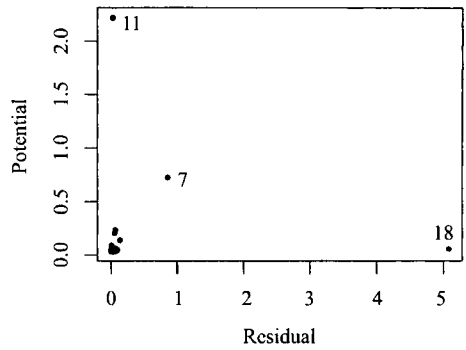

Figure 4.13 The Scottish Hills Races Data: Potential-Residual plot.

4.13 EFFECTS OF AN ADDITIONAL PREDICTOR

We discuss in general terms the effect of introducing a new variable in a regression equation. Two questions should be addressed: (a) Is the regression coefficient of the new variable significant? and (b) Does the introduction of the new variable substantially change the regression coefficients of the variables already in the regression equation? When a new variable is introduced in a regression equation, four possibilities result, depending on the answer to each of the above questions:

- **Case A:** The new variable has an insignificant regression coefficient and the remaining regression coefficients do not change substantially from their previous values. Under these conditions the new variable should not be included in the regression equation, unless some other external conditions (e.g., theory or subject matter considerations) dictate its inclusion.

- **Case B:** The new variable has a significant regression coefficient, and the regression coefficients for the previously introduced variables are changed in a substantial way. In this case the new variable should be retained, but an examination of collinearity[1] should be carried out. If there is no evidence of collinearity, the variable should be included in the equation and other additional variables should be examined for possible inclusion. On the other hand, if the variables show collinearity, corrective actions, as outlined in Chapter 10 should be taken.

- **Case C:** The new variable has a significant regression coefficient, and the coefficients of the previously introduced variables do not change in any substantial way. This is the ideal situation and arises when the new variable is

[1]Collinearity occurs when the predictor variables are highly correlated. This problem is discussed in Chapters 9 and 10.

uncorrelated with the previously introduced variables. Under these conditions the new variable should be retained in the equation.

- **Case D:** The new variable has an insignificant regression coefficient, but the regression coefficients of the previously introduced variables are substantially changed as a result of the introduction of the new variable. This is a clear evidence of collinearity, and corrective actions have to be taken before the question of the inclusion or exclusion of the new variable in the regression equation can be resolved.

It is apparent from this discussion that the effect a variable has on the regression equation determines its suitability for being included in the fitted equation. The results presented in this chapter influence the formulation of different strategies devised for variable selection. Variable selection procedures are presented in Chapter 11.

4.14 ROBUST REGRESSION

Another approach (not discussed here), useful for the identification of outliers and influential observations, is *robust regression*; a method of fitting that gives less weight to points of high leverage. There is a vast amount of literature on robust regression. The interested reader is referred, for example, to the books by Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Staudte and Sheather (1990), Birkes and Dodge (1993). We must also mention the papers by Krasker and Welsch (1982), Coakley and Hettmansperger (1993), Chatterjee and Mächler (1997), and Billor, Chatterjee, and Hadi (2006), which incorporate ideas of bounding influence and leverage in fitting. In Section 13.5 we give a brief discussion of robust regression and present a numerical algorithm for robust fitting. Two examples are given as illustration.

EXERCISES

4.1 Check to see whether or not the standard regression assumptions are valid for each of the following data sets:

(a) The Milk Production data described in Section 1.3.1.

(b) The Right-To-Work Laws data described in Section 1.3.2 and given in Table 1.3.

(c) The Egyptian Skulls data described in Section 1.3.3.

(d) The Domestic Immigration data described in Section 1.3.4.

(e) The New York Rivers data described in Section 1.3.5 and given in Table 1.9.

4.2 Find a data set where regression analysis can be used to answer a question of interest. Then:

Table 4.6 Expanded Computer Repair Times Data: Length of Service Calls (Minutes) and Number of Units Repaired (Units)

Row	Units	Minutes	Row	Units	Minutes
1	1	23	13	10	154
2	2	29	14	10	166
3	3	49	15	11	162
4	4	64	16	11	174
5	4	74	17	12	180
6	5	87	18	12	176
7	6	96	19	14	179
8	6	97	20	16	193
9	7	109	21	17	193
10	8	119	22	18	195
11	9	149	23	18	198
12	9	145	24	20	205

(a) Check to see whether or not the usual multiple regression assumptions are valid.

(b) Analyze the data using the regression methods presented thus far, and answer the question of interest.

4.3 Consider the computer repair problem discussed in Section 2.3. In a second sampling period, 10 more observations on the variables Minutes and Units were obtained. Since all observations were collected by the same method from a fixed environment, all 24 observations were pooled to form one data set. The data appear in Table 4.6.

(a) Fit a linear regression model relating Minutes to Units.

(b) Check each of the standard regression assumptions and indicate which assumption(s) seems to be violated.

4.4 In an attempt to find unusual points in a regression data set, a data analyst examines the P-R plot (shown in Figure 4.14). Classify each of the unusual points on this plot according to type.

4.5 Name one or more graphs that can be used to validate each of the following assumptions. For each graph, sketch an example where the corresponding assumption is valid and an example where the assumption is clearly invalid.

(a) There is a linear relationship between the response and predictor variables.

(b) The observations are independent of each other.

(c) The error terms have constant variance.

(d) The error terms are uncorrelated.

(e) The error terms are normally distributed.

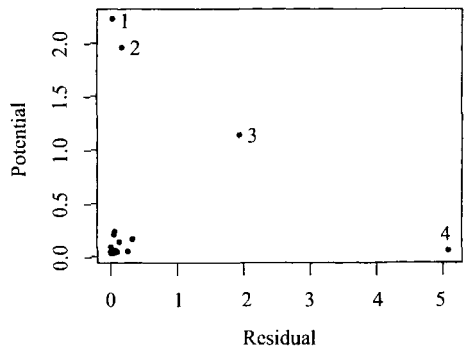

Figure 4.14 P-R plot used in Exercise 4.4.

(f) The observations are equally influential on least squares results.

4.6 The following graphs are used to verify some of the assumptions of the ordinary least squares regression of Y on $X_1, X_2, \ldots, X_p$:

1. The scatter plot of Y versus each predictor X_j.
2. The scatter plot matrix of the variables $X_1, X_2, \ldots, X_p$.
3. The normal probability plot of the internally standardized residuals.
4. The residuals versus fitted values.
5. The potential-residual plot.
6. Index plot of Cook's distance.
7. Index plot of Hadi's influence measure.

For each of these graphs:

(a) What assumption can be verified by the graph?

(b) Draw an example of the graph where the assumption does not seem to be violated.

(c) Draw an example of the graph which indicates the violation of the assumption.

4.7 Consider again the Cigarette Consumption data described in Exercise 3.14 and given in Table 3.17.

(a) What would you expect the relationship between Sales and each of the other explanatory variables to be (i.e., positive, negative)? Explain.

(b) Compute the pairwise correlation coefficients matrix and construct the corresponding scatter plot matrix.

(c) Are there any disagreements between the pairwise correlation coefficients and the corresponding scatter plot matrix?

(d) Is there any difference between your expectations in part (a) and what you see in the pairwise correlation coefficients matrix and the corresponding scatter plot matrix?

(e) Regress Sales on the six predictor variables. Is there any difference between your expectations in part (a) and what you see in the regression coefficients of the predictor variables? Explain inconsistencies if any.

(f) How would you explain the difference in the regression coefficients and the pairwise correlation coefficients between Sales and each of the six predictor variables?

(g) Is there anything wrong with the tests you made and the conclusions you reached in Exercise 3.14?

4.8 Consider again the Examination Data used in Exercise 3.3 and given in Table 3.10:

(a) For each of the three models, draw the P-R plot. Identify all unusual observations (by number) and classify as outlier, high leverage point, and/or influential observation.

(b) What model would you use to predict the final score F?

4.9 Either prove each of the following statements mathematically or demonstrate its correctness numerically using the Cigarette Consumption data described in Exercise 3.14 and given in Table 3.17:

(a) The sum of the ordinary least squares residuals is zero.

(b) The relationship between $\hat{\sigma}^2$ and $\hat{\sigma}^2_{(i)}$ is

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \left[\frac{n-p-1-r_i^2}{n-p-2} \right]. \tag{4.26}$$

4.10 Identify unusual observations for the data set in Table 4.7

Table 4.7 Data for Exercise 4.10

Row	Y	X	Row	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

4.11 Consider the Scottish Hills Races data in Table 4.5. Choose an observation index i (e.g., $i = 33$, which corresponds to the outlying observation number 33) and create an indicator (dummy) variable U_i, where all the values of U_i are zero except for its ith value which is one. Now consider comparing the following models:

$$H_0 : Time = \beta_0 + \beta_1\, Distance + \beta_2\, Climb + \varepsilon, \tag{4.27}$$

$$H_1 : Time = \beta_0 + \beta_1\, Distance + \beta_2\, Climb + \beta_3 U_i + \varepsilon. \quad (4.28)$$

Let r_i^* be the ith externally standardized residual obtained from fitting model (4.27). Show (or verify using an example) that

(a) The t-test for testing $\beta_3 = 0$ in Model (4.28) is the same as the ith externally standardized residual obtained from Model (4.27), that is, $t_3 = r_i^*$.

(b) The F-test for testing Model (4.27) versus (4.28) reduces to the square of the ith externally standardized residual, that is, $F = r_i^{*2}$.

(c) Fit Model (4.27) to the Scottish Hills Races data without the ith observation.

(d) Show that the estimates of β_0, β_1, and β_2 in Model (4.28) are the same as those obtained in c. Hence adding an indicator variable for the ith observation is equivalent to deleting the corresponding observation!

4.12 Consider the data in Table 4.8, which consist of a response variable Y and six predictor variables. The data can be obtained from the book's Web site. Consider fitting a linear model relating Y to all six X-variables.

(a) What least squares assumptions (if any) seem to be violated?

(b) Compute r_i, C_i, DFITS$_i$, and H_i.

(c) Construct the index plots of r_i, C_i, DFITS$_i$, and H_i as well as the Potential-Residual plot.

(d) Identify all unusual observations in the data and classify each according to type (i.e., outliers, leverage points, etc.).

4.13 Consider again the data set in Table 4.8. Suppose now that we fit a linear model relating Y to the first three X-variables. Justify your answer to each of the following questions with the appropriate added-variable plot:

(a) Should we add X_4 to the above model? If yes, keep X_4 in the model.

(b) Should we add X_5 to the above model? If yes, keep X_5 in the model.

(c) Should we add X_6 to the above model?

(d) Which model(s) would you recommend as the best possible description of Y? Use the above results and/or perform additional analysis if needed.

4.14 Consider fitting the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, to the data set in Table 4.8. Now let u be the residuals obtained from regressing Y on X_1. Also, let X_2 and v be the residuals obtained from regressing X_3 on X_1. Show (or verify using the data set in Table 4.8 as an example) that:

(a) $\hat{\beta}_3 = \sum_{i=1}^{n} u_i v_i / \sum_{i=1}^{n} v_i^2$

(b) The standard error of $\hat{\beta}_3$ is $\hat{\sigma} / \sqrt{\sum_{i=1}^{n} v_i^2}$.

Table 4.8 Data for Exercises 4.12–4.14

Row	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	443	49	79	76	8	15	205
2	290	27	70	31	6	6	129
3	676	115	92	130	0	9	339
4	536	92	62	92	5	8	247
5	481	67	42	94	16	3	202
6	296	31	54	34	14	11	119
7	453	105	60	47	5	10	212
8	617	114	85	84	17	20	285
9	514	98	72	71	12	−1	242
10	400	15	59	99	15	11	174
11	473	62	62	81	9	1	207
12	157	25	11	7	9	9	45
13	440	45	65	84	19	13	195
14	480	92	75	63	9	20	232
15	316	27	26	82	4	17	134
16	530	111	52	93	11	13	256
17	610	78	102	84	5	7	266
18	617	106	87	82	18	7	276
19	600	97	98	71	12	8	266
20	480	67	65	62	13	12	196
21	279	38	26	44	10	8	110
22	446	56	32	99	16	8	188
23	450	54	100	50	11	15	205
24	335	53	55	60	8	0	170
25	459	61	53	79	6	5	193
26	630	60	108	104	17	8	273
27	483	83	78	71	11	8	233
28	617	74	125	66	16	4	265
29	605	89	121	71	8	8	283
30	388	64	30	81	10	10	176
31	351	34	44	65	7	9	143
32	366	71	34	56	8	9	162
33	493	88	30	87	13	0	207
34	648	112	105	123	5	12	340
35	449	57	69	72	5	4	200
36	340	61	35	55	13	0	152
37	292	29	45	47	13	13	123
38	688	82	105	81	20	9	268
39	408	80	55	61	11	1	197
40	461	82	88	54	14	7	225

Source: Chatterjee and Hadi (1988)