Chapter 13
Model Diagnostics

This book deals with linear model theory, and as such we have assumed that the data are good and that the models are true. Unfortunately, good data are rare and true models are even rarer. Chapters 13 through 15 discuss some additional tools used by statisticians to deal with the problems presented by real data.

All models are based on assumptions. We typically assume that $\mathrm{E}(Y)$ has a linear structure, that the observations are independent, that the variance is the same for each observation, and that the observations are normally distributed. In truth, these assumptions will probably never be correct. It is our hope that if we check the assumptions and if the assumptions look plausible, then the mathematical methods presented here will work quite well.

If the assumptions are checked and found to be implausible, we need to have alternate ways of analyzing the data. In Section 2.7, Section 3.8, Chapter 10, and Chapter 12, we discussed the analysis of linear models with general covariance matrices. If an approximate covariance matrix can be found, the methods presented earlier can be used. (See also the discussion of the deleterious effects of estimating covariance matrices in Christensen (2001, Section 6.5).) Another approach is to find a transformation of the data so that the assumptions seem plausible for a standard linear model in the transformed data.

The primary purpose of this chapter is to present methods of identifying when there may be trouble with the assumptions. Analysis of the residuals is the method most often used for detecting invalidity of the assumptions. Residuals are used to check for nonnormality of errors, nonindependence, lack of fit, heteroscedasticity (inequality) of variances, and outliers (unusual data). They also help identify influential observations.

The vector of residuals is, essentially, an estimate of e in the model

$$Y = X\beta + e, \quad \mathrm{E}(e) = 0, \quad \mathrm{Cov}(e) = \sigma^2 I.$$

The residual vector is

$$\hat{e} = Y - X\hat{\beta} = (I - M)Y,$$

with

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*,
Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_13,

$$\mathrm{E}(\hat{e}) = (I-M)X\beta = 0$$

and

$$\mathrm{Cov}(\hat{e}) = (I-M)\sigma^2 I(I-M)' = \sigma^2(I-M).$$

For many of the techniques that we will discuss, the residuals are standardized so that their variances are about 1. The *standardized residuals* are

$$r_i = \hat{e}_i \Big/ \sqrt{MSE(1-m_{ii})},$$

where m_{ii} is the ith diagonal element of M and $\hat{e} = [\hat{e}_1, \ldots, \hat{e}_n]'$.

When checking for nonnormality or heteroscedastic variances, it is important to use the standardized residuals rather than unstandardized residuals. As just seen, the ordinary residuals have heteroscedastic variances. Before they are useful in checking for equality of the variances of the observations, they need to be standardized. Moreover, methods for detecting nonnormality are often sensitive to inequality of variances, so the use of ordinary residuals can make it appear that the errors are not normal even when they are.

Some computer programs used to use $\hat{e}/\sqrt{MSE}$ as standardized residuals, but I hope that is a thing of the past. This method is inferior to the standardization given above because it ignores the fact that the variances of the residuals are not all equal. The standardized residuals are also sometimes called the *Studentized residuals*, but that term is also sometimes used for another quantity discussed in Section 6.

Influential observations have been mentioned. What are they? One idea is that an observation is influential if it greatly affects the fitted regression equation. Influential observations are not intrinsically good or bad, but they are always important. Typically, influential observations are outliers: data points that are, in some sense, far from the other data points being analyzed. This happens in two ways. First, the y value associated with a particular row of the X matrix can be unlike what would be expected from examining the rest of the data. Second, a particular row of the X matrix can be unlike any of the other rows of the X matrix.

A frequently used method for analyzing residuals is to plot the (standardized) residuals against various other variables. We now consider an example that will be used throughout this chapter to illustrate the use of residual plots. In the example, we consider a model that will later be perturbed in various ways. This model and its perturbations will provide residuals that can be plotted to show characteristics of residual plots and the effects of unsatisfied assumptions.

EXAMPLE 13.0.1. Draper and Smith (1998) presented an example with 25 observations on a dependent variable, pounds of steam used by a company per month, and two predictor variables: x_1, the average atmospheric temperature for the month (in °F); and x_2, the number of operating days in the month. The values of x_1 and x_2 are listed in Table 13.1.

Draper and Smith's fitted equation is

$$y = 9.1266 - 0.0724x_1 + 0.2029x_2.$$

Table 13.1 Steam Data

Obs. no.	x_1	x_2	Obs. no.	x_1	x_2
1	35.3	20	14	39.1	19
2	29.7	20	15	46.8	23
3	30.8	23	16	48.5	20
4	58.8	20	17	59.3	22
5	61.4	21	18	70.0	22
6	71.3	22	19	70.0	11
7	74.4	11	20	74.5	23
8	76.7	23	21	72.1	20
9	70.7	21	22	58.1	21
10	57.5	20	23	44.6	20
11	46.4	20	24	33.4	20
12	28.9	21	25	28.6	22
13	28.1	21			

Our examples will frequently be set up so that the *true* model is

$$y_i = 9.1266 - 0.0724x_{i1} + 0.2029x_{i2} + e_i \,. \tag{1}$$

The vector $Y = (y_1, \ldots, y_{25})'$ can be obtained by generating the e_is and adding the terms on the right-hand side of the equation. Once Y is obtained, the equation $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ can be fitted by least squares and the residuals computed. By generating errors that have independent identical normal distributions, nonnormal distributions, serial correlations, or unequal variances, we can examine how residual plots should look when these conditions exist. By fitting models with incorrect mean structure, we can examine how residual plots for detecting lack of fit should look.

The material in this chapter is presented with applications to regression models in mind, but can be applied to ANOVA models with little modification. Excellent discussions of residuals and influential observations are found in Cook and Weisberg (1982), Atkinson (1985), and elsewhere. Cook and Weisberg (1994) give an extensive discussion of regression graphics.

Exercise 13.1 Show that $\hat{e}$ is the BLUP of e.

13.1 Leverage

A data point (case) that corresponds to a row of the X matrix that is "unlike" the other rows is said to have high leverage. In this section we will define the Mahalanobis distance and use it as a basis for identifying rows of X that are unusual. It will be shown that for regression models that include an intercept (a column of 1s),

the diagonal elements of the ppo M and the Mahalanobis distances are equivalent measures. In particular, a diagonal element of the ppo is an increasing function of the corresponding Mahalanobis distance. (There are some minor technical difficulties with this claim when X is not a full rank matrix.) This equivalence justifies the use of the diagonal elements to measure the abnormality of a row of the model matrix. *The diagonal elements of the ppo are the standard tool used for measuring leverage*.

In addition to their interpretation as a measure of abnormality, it will be shown that the diagonal elements of the projection matrix can have direct implications on the fit of a regression model. A diagonal element of the projection matrix that happens to be near 1 (the maximum possible) will force the estimated regression equation to go very near the corresponding y value. Thus, cases with extremely large diagonal elements have considerable influence on the estimated regression equation. It will be shown through examples that diagonal elements that are large, but not near 1, can also have substantial influence.

High leverage points are not necessarily bad. If a case with high leverage is consistent with the remainder of the data, then the case with high leverage causes no problems. In fact, the case with high leverage can greatly reduce the variability of the least squares fit. In other words, with an essentially correct model and good data, high leverage points actually help the analysis.

On the other hand, high leverage points are dangerous. The regression model that one chooses is rarely the true model. Usually it is only an approximation of the truth. High leverage points can change a good approximate model into a bad approximate model. An approximate model is likely to work well only on data that are limited to some particular range of values. It is unreasonable to expect to find a model that works well in places where very little data were collected. By definition, high leverage points exist where very little data were collected, so one would not expect them to be modeled well. Ironically, just the opposite result usually occurs. The high leverage points are often fit very well, while the fit of the other data is often harmed. The model for the bulk of the data is easily distorted to accommodate the high leverage points. When high leverage points are identified, the researcher is often left to decide between a bad model for the entire data and a good model for a more limited problem.

The purpose of this discussion of cases with high leverage is to make one point. If some data were collected in unusual places, then the appropriate goal may be to find a good approximate model for the area in which the bulk of the data were collected. This is not to say that high leverage points should always be thrown out of a data set. High leverage points need to be handled with care, and the implications of excluding high leverage points from a particular data set need to be thoroughly examined. High leverage points can be the most important cases in the entire data.

We begin by defining the Mahalanobis distance and establishing its equivalence to the diagonal elements of the projection operator. This will be followed by an examination of diagonal elements that are near 1. The section closes with a series of examples.

13.1.1 Mahalanobis Distances

The *Mahalanobis distance* measures how far a random vector is from the middle of its distribution. For this purpose, we will think of the rows of the matrix X as a sample of vectors from some population. Although this contradicts our assumption that the matrix X is fixed and known, our only purpose is to arrive at a reasonable summary measure of the distance of each row from the other rows. The Mahalanobis distance provides such a measure. The notation and ideas involved in estimating Mahalanobis distances are similar to those used in estimating best linear predictors. Estimation of best linear predictors was discussed in Subsection 6.3.4. In particular, we write the ith row of X as $(1, x_i')$ so that the corresponding linear model contains an intercept.

Let x be a random vector.

Definition 13.1.1. Let $\mathrm{E}(x) = \mu$ and $\mathrm{Cov}(x) = V$. The *squared Mahalanobis distance* is

$$D^2 = (x-\mu)'V^{-1}(x-\mu).$$

For a sample $x_1, \ldots, x_n$, the relative distances of the observations from the center of the distribution can be measured by the squared distances

$$D_i^2 = (x_i-\mu)'V^{-1}(x_i-\mu), \quad i = 1, \ldots, n.$$

Usually, μ and V are not available, so they must be estimated. Write

$$Z = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}.$$

Then μ can be estimated with $\bar{x}_\cdot' = (1/n)J_1^n Z$, and V can be estimated with

$$S = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i-\bar{x}_\cdot)(x_i-\bar{x}_\cdot)'\right] = \frac{1}{n-1}Z'\left(I-\frac{1}{n}J_n^n\right)Z.$$

Definition 13.1.2. The *estimated squared Mahalanobis distance* for the ith case in a sample of vectors $x_1, \ldots, x_n$ is

$$\hat{D}_i^2 = (x_i-\bar{x}_\cdot)'S^{-1}(x_i-\bar{x}_\cdot).$$

Note that the values of $\hat{D}^2$ are precisely the diagonal elements of

$$(n-1)\left(I-\frac{1}{n}J_n^n\right)Z\left[Z'\left(I-\frac{1}{n}J_n^n\right)Z\right]^{-1}Z'\left(I-\frac{1}{n}J_n^n\right).$$

Our interest in these definitions is that for a regression model $Y = X\beta + e$, the distance of the ith row of X from the other rows can be measured by the estimated squared Mahalanobis distance $\hat{D}_i^2$. In this context the rows of X are treated as a sample from some population. As mentioned earlier, when the model has an intercept, the diagonal elements of M and the estimated squared Mahalanobis distances are equivalent measures. When an intercept is included in the model, X can be written as $X = [J, Z]$. Since the rows of J are identical, the matrix S, defined for the entire matrix X, is singular. Thus, S^{-1} does not exist and the estimated Mahalanobis distances are not defined. Instead, we can measure the relative distances of the rows of Z from their center.

Theorem 13.1.3. Consider the linear model $Y = X\beta + e$, where X is a full rank matrix and $X = [J, Z]$. Then,

$$m_{ii} = \frac{1}{n} + \frac{\hat{D}^2}{n-1}.$$

PROOF. The theorem follows immediately from the fact that

$$M = \frac{1}{n}J_n^n + \left(I - \frac{1}{n}J_n^n\right) Z \left[Z'\left(I - \frac{1}{n}J_n^n\right) Z\right]^{-1} Z'\left(I - \frac{1}{n}J_n^n\right)$$

(cf. Sections 6.2, 9.1, and 9.2). The inverse in the second term of the right-hand side exists because X, and therefore $\left(I - \frac{1}{n}J_n^n\right) Z$, are full rank matrices. □

From this theorem, it is clear that the rows with the largest squared Mahalanobis distances are precisely the rows with the largest diagonal elements of the perpendicular projection matrix. For identifying high leverage cases in a regression model with an intercept, using the diagonal elements of M is equivalent to using the squared Mahalanobis distances.

Of course, for regression models that do not include an intercept, using the squared Mahalanobis distances is not equivalent to using the diagonal elements of the projection matrix. For such models it would probably be wise to examine both measures of leverage. The diagonal elements of the projection matrix are either given by or easily obtained from many computer programs. The information in the squared Mahalanobis distances can be obtained by the artifice of adding an intercept to the model and obtaining the diagonal elements of the projection matrix for the augmented model.

For linear models in which X is not of full rank, similar definitions could be made using a generalized inverse of the (estimated) covariance matrix rather than the inverse.

Exercise 13.2 Show that for a regression model that does not contain an intercept, the diagonal elements of the perpendicular projection operator are equivalent

to the estimated squared Mahalanobis distances computed with the assumption that $\mu = 0$.

13.1.2 Diagonal Elements of the Projection Operator

Having established that the diagonal elements of M are a reasonable measure of how unusual a row of X is, we are left with the problem of calibrating the measure. How big does m_{ii} have to be before we need to worry about it? The following proposition indirectly establishes several facts that allow us to provide some guidelines.

Proposition 13.1.4. For any i

$$m_{ii}(1 - m_{ii}) = \sum_{j \neq i} m_{ij}^2.$$

PROOF. Because M is a symmetric idempotent matrix,

$$m_{ii} = \sum_{j=1}^{n} m_{ij} m_{ji} = \sum_{j=1}^{n} m_{ij}^2 = m_{ii}^2 + \sum_{j \neq i} m_{ij}^2.$$

Subtracting gives

$$m_{ii}(1 - m_{ii}) = m_{ii} - m_{ii}^2 = \sum_{j \neq i} m_{ij}^2. \qquad \square$$

The term on the right-hand side of Proposition 13.1.4 is a sum of squared terms. This must be nonnegative, so $m_{ii}(1 - m_{ii}) \geq 0$. It follows immediately that the m_{ii}s must lie between 0 and 1.

Since the largest value that an m_{ii} can take is 1, any value near 1 indicates a point with extremely high leverage. Other values, considerably less than 1, can also indicate high leverage. Because $\text{tr}(M) = r(M)$, the average value of the m_{ii}s is p/n. Any m_{ii} value that is substantially larger than p/n indicates a point with high leverage. Some useful but imprecise terminology is set in the following definition.

Definition 13.1.5. Any case that corresponds to a row of the model matrix that is unlike the other rows is called an *outlier in the design space* (or estimation space). Any case corresponding to an m_{ii} substantially larger than p/n is called a case with *high leverage*. Any case corresponding to an m_{ii} near 1 is called a case with *extremely high leverage*.

Points with extremely high leverage have dramatic effects. If m_{ii} happens to be near 1, then $m_{ii}(1 - m_{ii})$ must be near zero and the right-hand side of Proposition 13.1.4 must be near zero. Since the right-hand side is a sum of squared terms,

for all $j \neq i$ the terms m_{ij} must be near zero. As will be shown, this causes a point with extremely high leverage to dominate the fitting process.

Let ρ_i be a vector of zeros with a 1 in the ith row, and let $x_i' = \rho_i' X$. The mean for the ith case is $\rho_i' X\beta = x_i'\beta$, which is estimated by $x_i'\hat{\beta} = \rho_i' MY = \sum_{j=1}^{n} m_{ij} y_j$. If m_{ii} is close to 1, m_{ij} is close to zero for all $j \neq i$; thus, the rough approximation $\rho_i' MY \doteq y_i$ applies. This approximation is by no means unusual. It simply says that the model fits well; however, the fact that the estimate largely ignores observations other than y_i indicates that something strange is occurring. Since m_{ii} is near 1, x_i' is far from the other rows of X. Thus, there is little information available about behavior at x_i' other than the observation y_i.

The fact that y_i is fit reasonably well has important implications for the estimated regression equation $f(x) = x'\hat{\beta}$. This function, when evaluated at x_i, must be near y_i. Regardless of whether y_i is an aberrant observation or whether a different approximate model is needed for observations taken near x_i, the estimated regression equation will adjust itself to fit y_i reasonably well. If necessary, the estimated regression equation will ignore the structure of the other data points in order to get a reasonable fit to the point with extremely high leverage. Thus, points with extremely high leverage have the potential to influence the estimated regression equation a great deal.

13.1.3 Examples

EXAMPLE 13.1.6. *Simple Linear Regression.*
Consider the model $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \ldots, 6$, $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, $x_5 = 5$, $x_6 = 15$. x_6 is far from the other x values, so it should be a case with high leverage. In particular, $m_{66} = .936$, so case six is an extremely high leverage point. (Section 6.1 gives a formula for M.)

For $i = 1, \ldots, 5$, data were generated using the model

$$y_i = 2 + 3x_i + e_i,$$

with the e_is independent $N(0,1)$ random variates. The y values actually obtained were $y_1 = 7.455$, $y_2 = 7.469$, $y_3 = 10.366$, $y_4 = 14.279$, $y_5 = 17.046$. The model was fit under three conditions: (1) with the sixth case deleted, (2) with $y_6 = 47 = 2 + 3(x_6)$, and (3) with $y_6 = 11 = 2 + 3(x_3)$. The results of fitting the simple linear regression model are summarized in the table below.

Condition	y_6	$\hat{y}_6$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\sqrt{MSE}$	dfE
1	deleted	43.91	3.425	2.699	1.423	3
2	47	46.80	2.753	2.937	1.293	4
3	11	13.11	10.599	0.1674	4.345	4

Figure 13.1 contains a scatterplot of the data under condition (3) and the fitted lines for conditions (1) and (3). Under conditions (1) and (2), reasonably consistent

fits are obtained. In particular, the extremely high leverage case has little effect on point estimation in these situations where the data are good and the model is true. (The high leverage case could have a large effect in decreasing the size of interval estimates.)

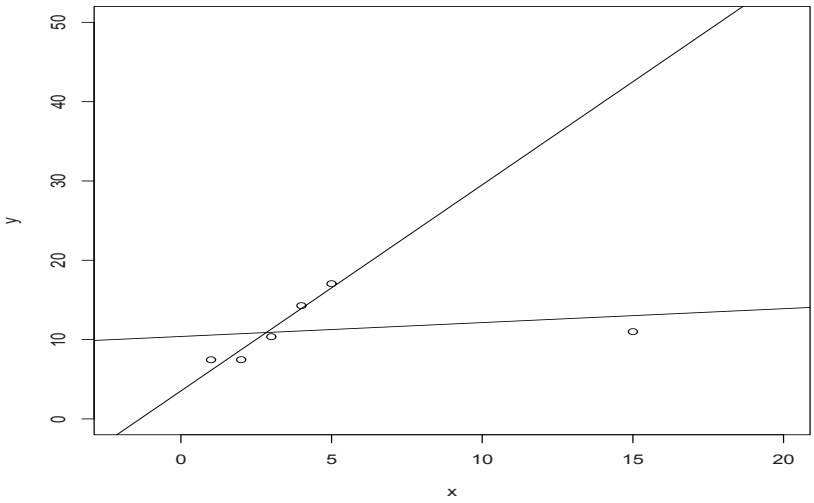

Fig. 13.1 Scatterplot and fitted lines for Example 13.1.6. (Note the optical illusions.)

Under condition (3), when the model is no longer valid for case six, the fit is grossly different from those obtained under conditions (1) and (2). When the extremely high leverage case is inconsistent with the rest of the data, the fit of the regression equation is dominated by that case. Note that the first five cases lead us to expect a value for y_6 in the mid-40s, but with $y_6 = 11$, the fitted value is 13.11, close to 11 and far from the mid-40s. Finally, the fit of the model, as measured by the MSE, is good under conditions (1) and (2), but much poorer under condition (3).

Other things being equal, under condition (2) it would be wise to include case six in the analysis. Under condition (3), it might be wiser not to try to model all the data, but rather to model only the first five cases and admit that little is known about behavior at x values substantially larger than 5. Unfortunately, in order to distinguish between conditions (2) and (3), the true model must be known, which in practice is not the case.

Finally, a word about residuals for case six. Under condition (2), the regression equation is right on the data point, $\hat{e}_6 = 0.2$, and $r_6 = 0.611$. On the other hand, for condition (3) there is a reasonable amount of error. The residual is $\hat{e}_6 = -2.11$. Compared to the other residuals (cf. Figure 13.1), $\hat{e}_6$ is not large, but neither is it

extremely small. The standardized residuals have a standard deviation of about 1, so the standardized residual for case six, $r_6 = -1.918$, is quite substantial.

Another useful tool is to look at the predicted residuals. For case six, this involves looking at the predicted value with case six deleted, 43.91, and comparing that to the observed value. For condition (2), the predicted residual is 3.09. For condition (3), the predicted residual is -32.91, which seems immense. To get a better handle on what these numbers mean, we need to standardize the predicted residuals. Predicted residuals are discussed in more detail in Sections 5 and 6. Proposition 13.6.1 gives a simple formula for computing standardized predicted residuals. Using this formula, the standardized predicted residual for case six under condition (2) is $t_6 = 0.556$, which is very reasonable. For condition (3), it is $t_6 = -5.86$, which is large enough to cause concern. (Section 6 explains why t_6 is not huge under condition (3).)

EXAMPLE 13.1.7. We now modify Example 13.1.6 by adding a second point far away from the bulk of the data. This is done in two ways. First, the extremely high leverage point is replicated with a second observation taken at $x = 15$. Second, a high leverage point is added that is smaller than the bulk of the data. In both cases, the y values for the first five cases remain unchanged.

With two observations at $x = 15$, $m_{66} = m_{77} = 0.48$. This is well above the value $p/n = 2/7 = 0.29$. In particular, the diagonal values for the other five cases are all less than 0.29.

To illustrate the effect of the high leverage points on the regression equation, conditions (2) and (3) of the previous example were combined. In other words, the two y values for $x = 15$ were taken as 11 and 47. The estimated regression equation becomes $\hat{y} = 6.783 + 1.5140x$. The slope of the line is about halfway between the slopes under conditions (2) and (3). More importantly, the predicted value for $x = 15$ is 29.49. The regression equation is being forced near the mean of y_6 and y_7. (The mean is 29.)

One of the salient points in this example is the effect on the root mean squared error. Under condition (2), where the high leverage point was consistent with the other data, the root mean squared error was 1.293. Under condition (3), where the high leverage point was grossly inconsistent with the other data, the root mean squared error was 4.293. In this case, with two high leverage points, one consistent with the bulk of the data and one not, the root mean squared error is 11.567. This drastic change is because, with two y values so far apart, one point almost has to be an outlier. Having an outlier in the ys at a high leverage point has a devastating effect on all estimates, especially the estimated error.

In the second illustration, x_7 was taken as -9. This was based on adding a second observation as far to the left of the bulk of the data as $x_6 = 15$ is to the right. The leverages are $m_{66} = m_{77} = 0.63$. Again, this value is well above $2/7$ and is far above the other diagonal values, which are around 0.15.

To illustrate the effect of high leverage on estimation, the y values were taken as $y_6 = y_7 = 11$. The estimated regression equation was $11.0906 + 0.0906x$. The root

mean squared error was 3.921. The t statistic for testing that the slope equaled zero was 0.40.

In essence, the data in the second illustration have been reduced to three points: a point $x = -9$ with a y value of 11, a point $x = 15$ with a y value of 11, and a point $x = 3$ (the mean of x_1 to x_5) with a y value of 11.523 (the mean of y_1 to y_5). Compared to the high leverage points at -9 and 15, the five points near 3 are essentially replicates.

Both of these scenarios illustrate situations where the leverages contain gaps. The first illustration has no points with leverages between 0.28 and 0.48. The second has no leverages between 0.16 and 0.63. Such gaps in the leverages indicate the the predictor variables contain clusters of observations that are separated from each other.

The final example of this section illustrates that high leverage points are model dependent. Since our measure of leverage is based on the perpendicular projection operator onto $C(X)$, it is not surprising that changing the model can affect the leverage of a case. The example below is a polynomial regression where a particular case does not have high leverage for fitting a line, but does have high leverage for fitting a parabola.

EXAMPLE 13.1.8. *Quadratic Regression.*
Consider fitting the models

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$
$$y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + e_i,$$

$i = 1, \ldots, 7$. The values of the x_is used were $x_1 = -10$, $x_2 = -9$, $x_3 = -8$, $x_4 = 0$, $x_5 = 8$, $x_6 = 9$, $x_7 = 10$. Note that the value of x_4 appears to be in the center of the data. For fitting a straight line, that appearance is correct. For fitting a line, the leverage of the fourth case is 0.14.

The model matrix for the quadratic model is

$$X = \begin{bmatrix} 1 & -10 & 100 \\ 1 & -9 & 81 \\ 1 & -8 & 64 \\ 1 & 0 & 0 \\ 1 & 8 & 64 \\ 1 & 9 & 81 \\ 1 & 10 & 100 \end{bmatrix}.$$

Note that the choice of the x_is makes the second column of x orthogonal to the other two columns. An orthonormal basis for $C(X)$ is easily obtained, and thus the diagonal elements of M are also easily obtained. The value of m_{44} is 0.84, which is quite large. From inspecting the third column of the model matrix, it is clear that the fourth case is unlike the rest of the data.

To make the example more specific, for $i \neq 4$ data were generated from the model

$$\begin{aligned} y_i &= 19.6 + 0.4x_i - 0.1x_i^2 + e_i \\ &= -0.1(x_i - 2)^2 + 20 + e_i, \end{aligned}$$

with the e_is independent $N(0,1)$ random variables. The values 0, 11.5, and 19.6 were used for y_4. These values were chosen to illustrate a variety of conditions. The value 19.6 is consistent with the model given above. In particular, $19.6 = \mathrm{E}(y_4)$. The value $11.5 = \mathrm{E}(y_2 + y_6)/2$ should give a fit that is nearly linear. The value 0 is simply a convenient choice that is likely to be smaller than any other observation. The Y vector obtained was

$$Y = (6.230, 8.275, 8.580, y_4, 16.249, 14.791, 14.024)'.$$

Figure 13.2 contains a scatterplot of the data that includes all three values for y_4 as well as a plot of the true regression curve.

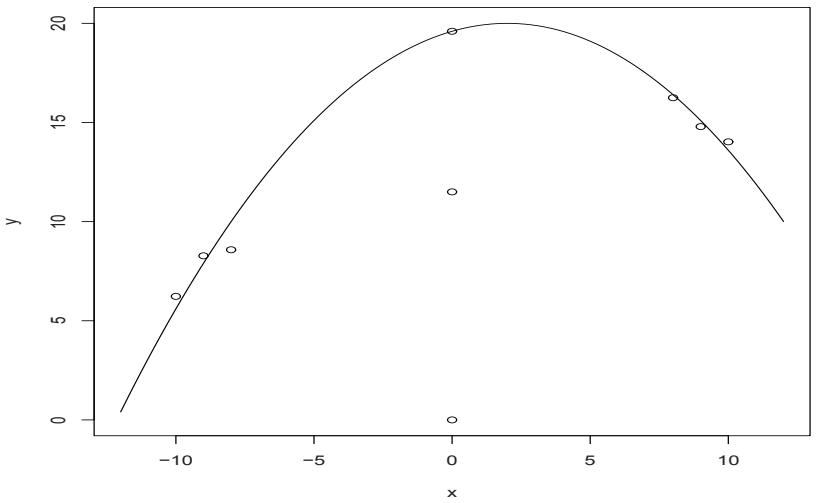

Fig. 13.2 Scatterplot for Example 13.1.8.

The linear and quadratic models were fitted with all three of the y_4 values and with the fourth case deleted from the model. For all models fitted, the coefficient of the linear term was 0.4040. As mentioned above, the second column (the linear column) of the matrix X is orthogonal to the other columns. Thus, for any value of y_4 the linear coefficient will be the same for the quadratic model and the linear model. The linear coefficient does not depend on the value of y_4 because $x_4 = 0$. Also, because $x_4 = 0$, the predicted value for y_4 is just the intercept of the line. Fitting simple linear regressions resulted in the following:

Linear Fits

y_4	$\hat{y}_4 = \hat{\beta}_0$	$\sqrt{MSE}$	dfE
deleted	11.36	1.263	4
0.0	9.74	4.836	5
11.5	11.38	1.131	5
19.6	12.54	3.594	5

As designed, the fits for y_4 deleted and $y_4 = 11.5$ are almost identical. The other values of y_4 serve merely to move the intercept up or down a bit. They do not move the line enough so that the predicted value $\hat{y}_4$ is close to the observed value y_4. The y_4 values of 0 and 19.6 do not fit the line well, which is reflected in the increased values for the root mean squared error. In summary, the values of y_4 do not have a great effect on the fitted lines.

The results of the quadratic fits, including the t statistic for testing $\gamma_2 = 0$, are

Quadratic Fits

y_4	$\hat{y}_4 = \hat{\gamma}_0$	$\hat{\gamma}_2$	$t(\gamma_2)$	$\sqrt{MSE}$	dfE
deleted	16.564	−0.064	−3.78	0.607	3
0.0	2.626	0.102	2.55	3.339	4
11.5	12.303	−0.013	−0.97	1.137	4
19.6	19.119	−0.094	−9.83	0.802	4

As expected, the y_4 deleted and $y_4 = 19.6$ situations are similar, and approximate the true model. The $y_4 = 11.5$ situation gives essentially a straight line; the t statistic for testing $H_0 : \gamma_2 = 0$ is very small. The true quadratic structure of all but one case is ignored in favor of the linear fit. (Note that the root mean squared error is almost identical in the linear and quadratic fits when $y_4 = 11.5$.) Finally, with $y_4 = 0$, the entire structure of the problem is turned upside down. The true model for all cases except case four is a parabola opening down. With $y_4 = 0$, the fitted parabola opens up. Although the fourth case does not have high leverage for the linear fits, the fourth case greatly affects the quadratic fits.

It is important to note that the problems caused by high leverage points are not unique to fitting models by least squares. When fitting by least squares, high leverage points are fit well, because it is assumed that the model is correct for all of the data points. Least squares accommodates all the data points, including the points with high leverage. If a model, say model A, fits the bulk of the data, but a different model, say model B, is necessary to explain the data when including the cases with high leverage, then the error of fitting model A to the high leverage cases is likely to be large. Any method of fitting models (e.g., robust regression) that seeks to minimize errors will modify the fit so that those large errors do not occur. Thus, any fitting mechanism forces the fitted model to do a reasonable job of fitting all of the data. Since the high leverage cases must be fit reasonably well and since, by definition, data are sparse near the high leverage points, the high leverage points are often fit extremely well.

13.2 Checking Normality

We give a general discussion of the problem of checking normality for a random sample and then relate it to the analysis of residuals. Suppose $v_1, \ldots, v_n$ are i.i.d. $N(\mu, \sigma^2)$ and $z_1, \ldots, z_n$ are i.i.d. $N(0,1)$. Ordering these from smallest to largest gives the order statistics $v_{(1)} \le \cdots \le v_{(n)}$ and $z_{(1)} \le \cdots \le z_{(n)}$. The expected values of the standard normal order statistics are $\mathrm{E}[z_{(1)}], \ldots, \mathrm{E}[z_{(n)}]$. Since the v_is are normal, $[v_{(i)} - \mu]/\sigma \sim z_{(i)}$ and we should have the approximate equality, $[v_{(i)} - \mu]/\sigma \doteq \mathrm{E}[z_{(i)}]$ or $v_{(i)} \doteq \sigma \mathrm{E}[z_{(i)}] + \mu$.

Suppose now that $v_1, \ldots, v_n$ are observed and we want to see if they are a random sample from a normal distribution. If the v_is are from a normal distribution, a graph of the pairs $(\mathrm{E}[z_{(i)}], v_{(i)})$ should be an approximate straight line. If the graph is not an approximate straight line, nonnormality is indicated. These graphs are variously called *rankit plots*, *normal plots*, or *q-q plots*.

To make the graph, one needs the values $\mathrm{E}[z_{(i)}]$. These values, often called *rankits*, *normal scores*, or *theoretical (normal) quantiles*, are frequently approximated as follows. Let

$$\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp[-t^2/2] dt.$$

$\Phi(x)$ is the cumulative distribution function for a standard normal random variable. Let u have a uniform distribution on the interval $(0,1)$. Write $u \sim U(0,1)$. It can be shown that

$$\Phi^{-1}(u) \sim N(0,1).$$

If $z_1, \ldots, z_n$ are i.i.d. $N(0,1)$ and $u_1, \ldots, u_n$ are i.i.d. $U(0,1)$, then

$$z_{(i)} \sim \Phi^{-1}(u_{(i)}),$$

and

$$\mathrm{E}[z_{(i)}] = \mathrm{E}[\Phi^{-1}(u_{(i)})].$$

One reasonable approximation for $\mathrm{E}[z_{(i)}]$ is

$$\mathrm{E}[z_{(i)}] \doteq \Phi^{-1}(\mathrm{E}[u_{(i)}]) = \Phi^{-1}\left(\frac{i}{n+1}\right).$$

In practice, better approximations are available. Take

$$\mathrm{E}[z_{(i)}] \doteq \Phi^{-1}\left(\frac{i-a}{n+(1-2a)}\right). \tag{1}$$

For $n \ge 5$, an excellent approximation is $a = 3/8$, see Blom (1958). The R programming language defaults to $a = 3/8$ when $n \le 10$ with $a = 0.5$ otherwise. MINITAB always uses $a = 3/8$.

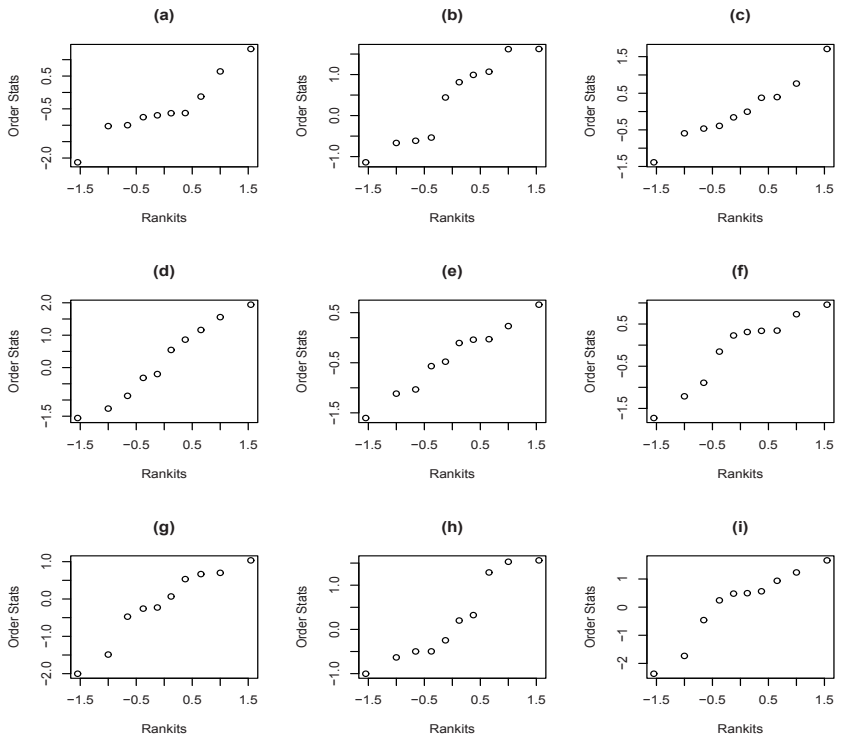

Fig. 13.3 Normal plots for normal data, $n = 10$.

To check whether $e \sim N(0, \sigma^2 I)$ in a linear model, the standardized residuals are plotted against the rankits. If the plot is not linear, nonnormality is suspected.

EXAMPLE 13.2.1. For $n = 10, 25, 50$, nine random vectors, say Ei, $i = 1, \ldots, 9$, were generated so that the Eis were independent and

$$Ei \sim N(0, I).$$

The corresponding Ei — rankit plots in Figures 13.3 through 13.5 give an idea of how straight one can reasonably expect rankit plots to be for normal data. All plots use rankits from equation (1) with $a = 3/8$.

Figure 13.3 gives rankit plots for random samples of size $n = 10$. Notice the substantial deviations from linearity in these plots even though the data are iid normal. Figure 13.4 gives rankit plots for samples of size $n = 25$ and Figure 13.5 gives plots for samples of size $n = 50$. As the sample sizes increase, the plots become more linear.

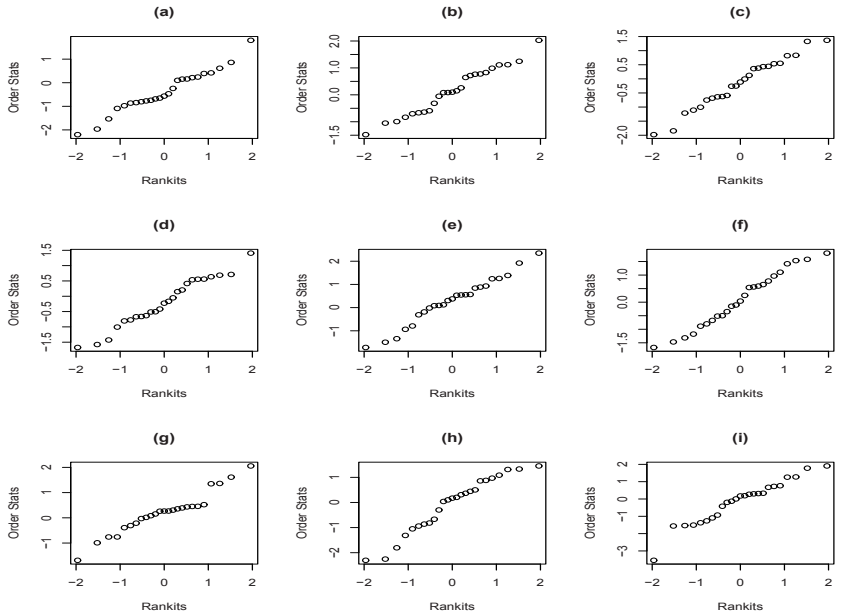

Fig. 13.4 Normal plots for normal data, $n = 25$.

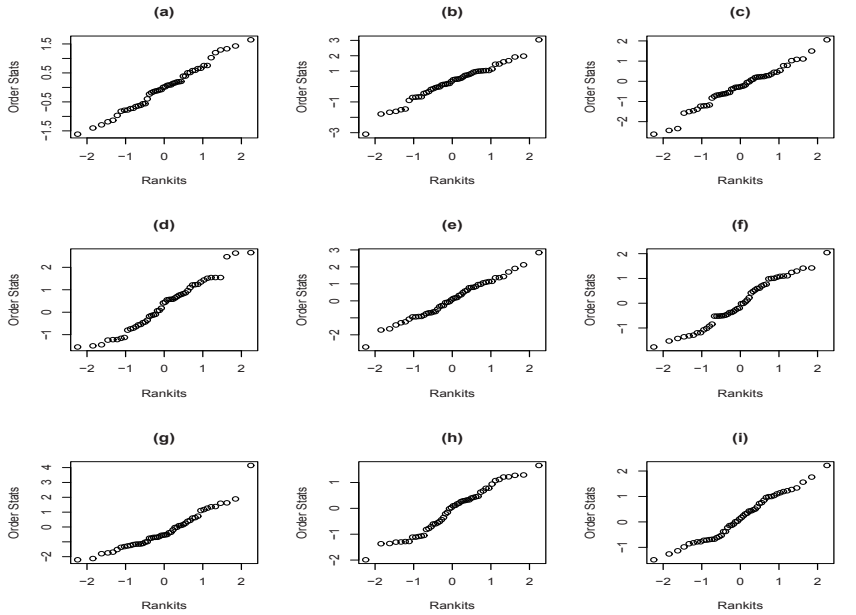

Fig. 13.5 Normal plots for normal data, $n = 50$.

In addition, for $n = 25$ the Eis were used with (13.0.1) to generate nine Y vectors and the standardized residuals from fitting

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \tag{2}$$

were computed. The Eis and the corresponding standardized residual vectors (Ris) were plotted against the approximate rankits. Two Ri — rankit plots are provided to give some idea of how the correlations among the residuals can affect the plots. Of the nine pairs of plots generated, only the two that look least normal (based on Ei as determined by the author) are displayed in Figures 13.6. The standardized residuals in plot (b) seem more normal than the original normal observations in plot (a).

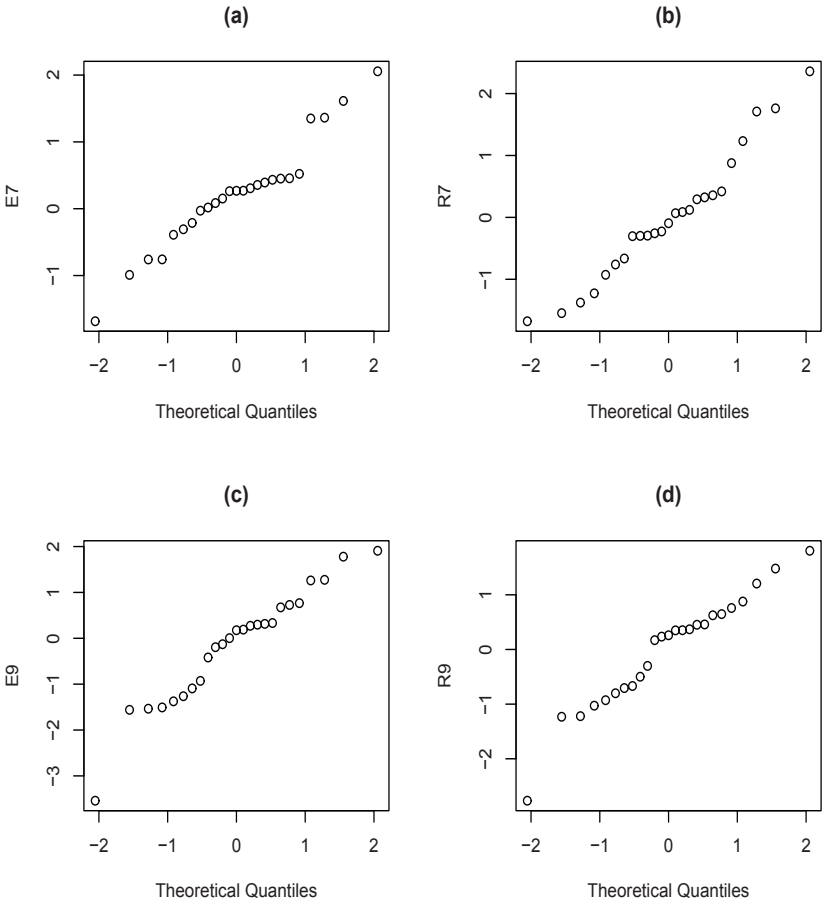

Fig. 13.6 Normal plots for normal data and corresponding standardized residual plots.

Duan (1981) established that the residuals provide an asymptotically consistent estimate of the underlying error distribution. Thus for large samples, the residual — rankit plot should provide an accurate evaluation of normality. In practice, however, the real question is not whether the data are nonnormal, but whether they are sufficiently nonnormal to invalidate a normal approximation. This is a more difficult question to address. See Arnold (1981, Chapter 10) for a discussion of *asymptotic consistency* of least squares estimates.

Shapiro and Wilk (1965) developed a formal test for normality related to normal plots. Unfortunately, the test involves knowing the inverse of the covariance matrix of $z_{(1)},\ldots,z_{(n)}$. An excellent approximation to their test was suggested by Shapiro and Francia (1972). The approximate test statistic is the square of the sample correlation coefficient computed from the pairs $(\mathrm{E}[z_{(i)}], v_{(i)})$, $i = 1,\ldots,n$. Let

$$W' = \left(\sum_{i=1}^{n} \mathrm{E}[z_{(i)}]v_{(i)}\right)^2 \Big/ \sum_{i=1}^{n}(\mathrm{E}[z_{(i)}])^2 \sum_{i=1}^{n}(v_{(i)} - \bar{v})^2.$$

(Note: $\sum_{i=1}^{n} \mathrm{E}[z_{(i)}] = 0$ by symmetry.) If W' is large, there is no evidence of nonnormality. Small values of W' are inconsistent with the hypothesis that the data are a random sample from a normal distribution. Approximate percentage points for the distribution of W' are given by Christensen (1996a).

To test for normality in linear models, the v_is are generally replaced by the r_is, i.e., the standardized residuals.

EXAMPLE 13.2.2. The W' statistics were computed for the Eis and Ris from Example 13.2.1 with $n = 25$. The values are listed below.

i	$W'(E)$	$W'(R)$
1	0.966	0.951
2	0.975	0.982
3	0.980	0.980
4	0.973	0.968
5	0.978	0.973
6	0.981	0.975
7	0.945	0.964
8	0.955	0.947
9	0.946	0.948

Note that, in this example (as in previous editions of the book), the W' values do not seem to be either systematically higher or lower when computed on the residuals rather than the true errors. Denoting the lower α percentile of W' from a sample of size n as $W'(\alpha,n)$, a simulation similar to that used in Christensen (1996a) gives $W'(0.01,25) = 0.874$ and $W'(0.05,25) = 0.918$. None of the tests are rejected.

To give some indication of the power of the W' test, the example was repeated using ten samples of data generated using nonnormal errors. In one case, the errors

were generated from a Cauchy distribution (a t distribution with 1 degree of freedom), and in a second case the errors were generated from a t distribution with three degrees of freedom. The results follow.

	Cauchy			$t(3)$	
i	$W'(E)$	$W'(R)$	i	$W'(E)$	$W'(R)$
1	0.491	0.553	1	0.861	0.871
2	0.539	0.561	2	0.878	0.966
3	0.903	0.909	3	0.891	0.856
4	0.822	0.783	4	0.654	0.637
5	0.575	0.644	5	0.953	0.951
6	0.354	0.442	6	0.912	0.905
7	0.502	0.748	7	0.978	0.979
8	0.753	0.792	8	0.958	0.959
9	0.921	0.952	9	0.896	0.881
10	0.276	0.293	10	0.972	0.967

With the Cauchy distribution, all but two of the tests are rejected at the 0.01 level, and one of these is rejected at the 0.05 level. With the $t(3)$ distribution, only two tests are rejected at the 0.01 level, with six of the tests based on E and five of the tests based on R rejected at 0.05.

The techniques for checking normality are applied directly to the standardized residuals. The theory assumed that the v_is were independent, but $\mathrm{Cov}(\hat{e}) = \sigma^2(I - M)$, so both the residuals and the standardized residuals are correlated. One way to avoid this problem is to consider the $(n-p) \times 1$ vector $O'\hat{e}$, where the columns of O are an orthonormal basis for $C(I-M)$.

$$\mathrm{Cov}(O'\hat{e}) = \sigma^2 O'(I-M)O = \sigma^2 O'O = \sigma^2 I,$$

so the procedures for checking normality can be validly applied to $O'\hat{e}$. The problem with this is that there are an infinite number of ways to pick O and the results depend on the choice of O. In fact, one can pick O so that $W' = 1$. Note that for any choice of O', O_1O' is another valid choice, where O_1 is any $n-p$ orthogonal matrix. Because O_1 is an arbitrary orthogonal matrix, $O_1O'\hat{e}$ can be any rotation of $O'\hat{e}$. In particular, it can be one that is in exactly the same direction as the vector $a_{n-p} = (a_{n-p,1}, \ldots, a_{n-p,n-p})'$, where $a_{n,i} = \mathrm{E}[z_{(i)}]$ from a sample of size n. The sample correlation between these two vectors will be 1; thus $W' = 1$.

Exercise 13.3 Show that the sample correlation between two (mean adjusted) vectors in the same direction is 1.

Exercise 13.4 Using the model of Example 13.2.2, estimate the power of detecting a $t(3)$ with $\alpha = 0.05$ by simulation.

13.2.1 Other Applications for Normal Plots

We close this section with two variations on the use of normal rankit plots.

EXAMPLE 13.2.3. Consider an ANOVA, say

$$y_{ijk} = \mu_{ij} + e_{ijk},$$

$i = 1,\ldots,a$, $j = 1,\ldots,b$, $k = 1,\ldots,N_{ij}$. Rather than using residuals to check whether the e_{ijk}s are i.i.d. $N(0,\sigma^2)$, for each pair i, j we can check whether $y_{ij1},\ldots,y_{ijN_{ij}}$ are i.i.d. $N(\mu_{ij},\sigma^2)$. The model assumes that for each treatment ij, the N_{ij} observations are a random sample from a normal population. Each treatment can be checked for normality individually. This leads to forming ab normal plots, one for each treatment. Of course, these plots will only work well if the N_{ij}s are reasonably large.

EXAMPLE 13.2.4. We now present a graphical method for evaluating the interaction in a two-way ANOVA with only one observation in each cell. The model is

$$y_{ij} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij} + e_{ij},$$

$i = 1,\ldots,a$, $j = 1,\ldots,b$. Here the e_{ij}s are assumed to be i.i.d. $N(0,\sigma^2)$. With one observation per cell, the $(\alpha\eta)_{ij}$s are confounded with the e_{ij}s; the effects of interaction cannot be separated from those of error.

As was mentioned in Section 7.2, the interactions in this model are often assumed to be nonexistent so that an analysis of the main effects can be performed. As an alternative to assuming no interaction, one can evaluate graphically an orthogonal set of $(a-1)(b-1)$ interaction contrasts, say $\lambda_{rs}'\beta$. If there are no interactions, the values $\lambda_{rs}'\hat{\beta}\Big/\sqrt{\lambda_{rs}'(X'X)^-\lambda_{rs}}$ are i.i.d. $N(0,\sigma^2)$. Recall that for an interaction contrast, $\lambda'\beta = \sum_{ij} q_{ij}(\alpha\eta)_{ij}$,

$$\lambda'\hat{\beta} = \sum_{ij} q_{ij} y_{ij}$$

and

$$\lambda'(X'X)^-\lambda = \sum_{ij} q_{ij}^2.$$

The graphical procedure is to order the $\lambda_{rs}'\hat{\beta}\Big/\sqrt{\lambda_{rs}'(X'X)^-\lambda_{rs}}$ values and form a normal rankit plot. If there are no interactions, the plot should be linear. Often there will be some estimated interactions that are near zero and some that are clearly nonzero. The near zero interactions should fall on a line, but clearly nonzero interactions will show up as deviations from the line. Contrasts that do not fit the line are identified as nonzero interaction contrasts (without having executed a formal test).

The interactions that fit on a line are used to estimate σ^2. This can be done in either of two ways. First, an estimate of the slope of the linear part of the graph can

be used as an estimate of the standard deviation σ. Second, sums of squares for the contrasts that fit the line can be averaged to obtain a mean squared error.

Both methods of estimating σ^2 are open to criticism. Consider the slope estimate and, in particular, assume that $(a-1)(b-1) = 12$ and that there are three nonzero contrasts all yielding large positive values of $\lambda_{rs}'\hat{\beta}/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$. In this case, the ninth largest value is plotted against the ninth largest rankit. Unfortunately, we do not know that the ninth largest value is the ninth largest observation in a random sample of size 12. If we could correct for the nonzero means of the three largest contrasts, what we observed as the ninth largest value could become anything from the ninth to the twelfth largest value. To estimate σ, we need to plot the mean adjusted statistics $\left(\lambda_{rs}'\hat{\beta} - \lambda_{rs}'\beta\right)/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$. We know that 9 of the 12 values $\lambda_{rs}'\beta$ are zero. The ninth largest value of $\lambda_{rs}'\hat{\beta}/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$ can be any of the order statistics of the mean adjusted values $\left(\lambda_{rs}'\hat{\beta} - \lambda_{rs}'\beta\right)/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$ between 9 and 12. The graphical method assumes that extreme values of $\lambda_{rs}'\hat{\beta}/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$ are also extreme values of the mean adjusted statistics. There is no justification for this assumption. If the ninth largest value were really the largest of the 12 mean adjusted statistics, then plotting the ninth largest value rather than the ninth largest mean adjusted value against the ninth largest rankit typically indicates a slope that is larger than σ. Thus the graphical procedure tends to overestimate the variance. Alternatively, the ninth largest value may not seem to fit the line and so, inappropriately, be declared nonzero. These problems should be ameliorated by dropping the three clearly nonzero contrasts and replotting the remaining contrasts as if they were a sample of size 9. In fact, the replotting method will tend to have a downward bias, as discussed in the next paragraph.

The criticism of the graphical procedure was based on what happens when there are nonzero interaction contrasts. The criticism of the mean squared error procedure is based on what happens when there are no nonzero interaction contrasts. In this case, if one erroneously identifies contrasts as being nonzero, the remaining contrasts have been selected for having small absolute values of $\lambda_{rs}'\hat{\beta}/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$ or, equivalently, for having small sums of squares. Averaging a group of sums of squares that were chosen to be small clearly underestimates σ^2. The author's inclination is to use the mean squared error criterion and try very hard to avoid erroneously identifying zero interactions as nonzero. This avoids the problem of estimating the slope of the normal plot. Simulation envelopes such as those discussed by Atkinson (1985, Chapter 4) can be very helpful in deciding which contrasts to identify as nonzero.

Some authors contend that, for visual reasons, normal plots should be replaced with plots that do not involve the sign of the contrasts, see Atkinson (1981, 1982). Rather than having a graphical procedure based on the values $\lambda_{rs}'\hat{\beta}/\sqrt{\lambda_{rs}'(X'X)^{-}\lambda_{rs}}$, the squared values, i.e., the sums of squares for the $\lambda_{rs}'\beta$ contrasts, can be used. When there are no interactions,

$$\frac{SS(\lambda_{rs}'\beta)}{\sigma^2} \sim \chi^2(1).$$

The contrasts are orthogonal so, with no interactions, the values $SS(\lambda'_{rs}\beta)$ form a random sample from a $\sigma^2\chi^2(1)$ distribution. Let $w_1,\ldots,w_r$ be i.i.d $\chi^2(1)$, where $r=(a-1)(b-1)$. Compute the expected order statistics $\mathrm{E}[w_{(i)}]$ and plot the pairs $\big(\mathrm{E}[w_{(i)}], SS(\lambda'_{(i)}\beta)\big)$, where $SS(\lambda'_{(i)}\beta)$ is the ith smallest of the sums of squares. With no interactions, this should be an approximate straight line through zero with slope σ^2. For nonzero contrasts, $SS(\lambda'_{rs}\beta)$ has a distribution that is σ^2 times a *noncentral* $\chi^2(1)$. Values of $SS(\lambda'_{rs}\beta)$ that are substantially above the linear portion of the graph indicate nonzero contrasts. A graphical estimate of σ^2 is available from the sums of squares that fit on a line; this has bias problems similar to that of a normal plot. The theoretical quantiles $\mathrm{E}[w_{(i)}]$ can be approximated by evaluating the inverse of the $\chi^2(1)$ cdf at $i/(n+1)$.

A corresponding method for estimating σ, based on the square roots of the sums of squares, is called a *half-normal plot*. The expected order statistics are often approximated as $\Phi^{-1}((n+i)/(2n+1))$.

These methods can be easily extended to handle other situations in which there is no estimate of error available. In fact, this graphical method was first proposed by Daniel (1959) for analyzing 2^n factorial designs. Daniel (1976) also contains a useful discussion.

13.3 Checking Independence

Lack of independence occurs when $\mathrm{Cov}(e)$ is not diagonal. One reason that good methods for evaluating independence are difficult to develop is that, unlike the other assumptions involved in $e\sim N(0,\sigma^2 I)$, independence is not a property of the population in question. Independence is a property of the way that the population is sampled. As a result, there is no way to check independence without thinking hard about the method of sampling. Identifying lack of independence is closely related to identifying lack of fit. For example, consider data from a randomized complete block (RCB) experiment being analyzed with a one-way analysis of variance model that ignores blocks. If the blocks have fixed effects, the one-way model suffers from lack of fit. If the blocks have random effects with a common mean, the one-way model suffers from lack of independence. We begin with a general discussion of ideas for testing the independence assumption based upon Christensen and Bedrick (1997). This is followed by a subsection on detecting serial correlation.

A key idea in checking independence is the formation of rational subgroups. To evaluate whether a group of numbers form a random sample from a population, Shewhart (1931) proposed using control charts for means. The means being charted were to be formed from rational subgroups of the observations that were obtained under essentially identical conditions. Shewhart (1939, p. 42) suggests that a control chart is less a test for whether data form a random sample and more an operational definition of what it means to have a random sample. It is easily seen that a means chart based on rational subgroups is sensitive to lack of independence, lack of fit (nonconstant mean), inequality of variances, and nonnormality. In analyzing linear

models, statisticians seek assurance that any lack of independence or other violations of the assumptions are not so bad as to invalidate their conclusions. Essentially, statisticians need an operational definition of when traditional linear model theory can be applied.

As used with linear models, rational subgroups are simply clusters of observations. They can be clustered in time, or in space, by having similar predictor variables, by being responses on the same individual, or by almost anything that the sampling scheme suggests could make observations within a cluster more alike than observations outside a cluster. To test for lack of independence, the near replicate lack of fit tests presented in Subsection 6.6.2 can be used. Simply replace the clusters of near replicates with clusters of rational subgroups determined by the sampling scheme. Christensen and Bedrick (1997) found that the analysis of covariance test, i.e., the Christensen (1989) test, worked well in a wide variety of situations, though the Shillington test often worked better when the clusters were very small. Of course, specialized tests for specific patterns of nonindependence can work much better than these general tests *when the specific patterns are appropriate*.

13.3.1 Serial Correlation

An interesting case of nonindependence is serial correlation. This occurs frequently when observations $y_1, y_2, \ldots, y_n$ are taken serially at equally spaced time periods. A model often used when the observations form such a time series is

$$\mathrm{Cov}(e) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix},$$

where $\rho_1, \rho_2, \ldots, \rho_{n-1}$ are such that $\mathrm{Cov}(e)$ is positive definite. Typically, only the first few of $\rho_1, \rho_2, \ldots, \rho_{n-1}$ will be substantially different from zero. One way of detecting serial correlation is to plot r_i versus i. If, say, ρ_1 and ρ_2 are positive and $\rho_3, \ldots, \rho_n$ are near zero, a plot of r_i versus i may have oscillations, but residuals that are adjacent should be close. If ρ_1 is negative, then ρ_2 must be positive. The plot of r_i versus i in this case may or may not show overall oscillations, but adjacent residuals should be oscillating rapidly. An effective way to detect a nonzero ρ_s is to plot (r_i, r_{i+s}) for $i = 1, \ldots, n-s$ or to compute the corresponding sample correlation coefficient from these pairs.

A special case of serial correlation is $\rho_s = \rho^s$ for some parameter ρ between -1 and 1. This AR(1), i.e., autoregressive order 1, covariance structure can be obtained by assuming 1) $e_1 \sim N(0, \sigma^2)$, and 2) for $i > 1$, $e_{i+1} = \rho e_i + v_{i+1}$, where $v_2, \ldots, v_n$ are i.i.d. $N(0, (1-\rho^2)\sigma^2)$ and v_{i+1} is independent of e_i for all i. Other models for

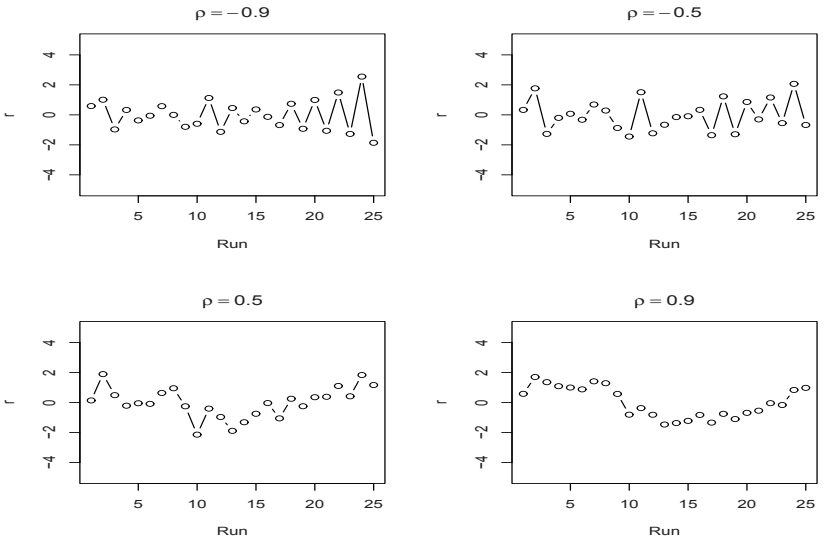

Fig. 13.7 Serial correlation standardized residual plots.

serial correlation are discussed in Christensen et al. (2010, Section 10.3). Most are based on ARMA time series models, cf. Christensen (2001, Chapter 5).

EXAMPLE 13.3.1. For ρ equal to -0.9, 0.5, 0.5, and 0.9, serially correlated error vectors were generated as just described. Dependent variable values y were obtained using (13.0.1) and the model (13.2.2) was fitted, giving a standardized residual vector r. For all values of ρ, the standardized residuals are plotted against their observation numbers. Within each figure, $z_1, z_2, \ldots, z_n$ are i.i.d. $N(0,1)$ with $e_1 = z_1$ and $v_i = \sqrt{1-\rho^2}z_i$, so only ρ changes. Figures 13.7 through 13.9 give three independent sets of plots. Note that when ρ is positive, adjacent observations remain near one another. The overall pattern tends to oscillate slowly. When ρ is negative, the observations oscillate very rapidly; adjacent observations tend to be far apart, but observations that are one apart (e.g., e_i and e_{i+2}) are fairly close.

Figures 13.10 through 13.14 contain plots with $\rho = 0$. Figure 13.10 is in the same form as Figures 13.7 through 13.9. The other figures use a different style. Comparing Figure 13.10 with Figures 13.7–13.9, it does not seem easy to distinguish between $\rho = 0$ and moderate correlations like $\rho = \pm 0.5$.

Figures 13.10 through 13.14 are of interest not only for illustrating a lack of serial correlation, but also as examples of what the plots in Section 4 should look like, i.e., these are standardized residual plots when all the model assumptions are valid. Note that the horizontal axis is not specified, because the residual plots should show no correlation, regardless of what they are plotted against. It is interesting to

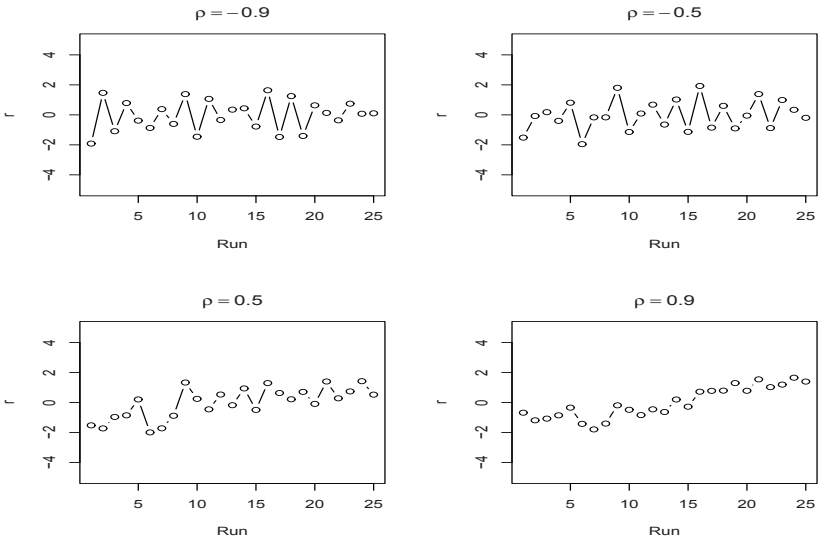

Fig. 13.8 Serial correlation standardized residual plots.

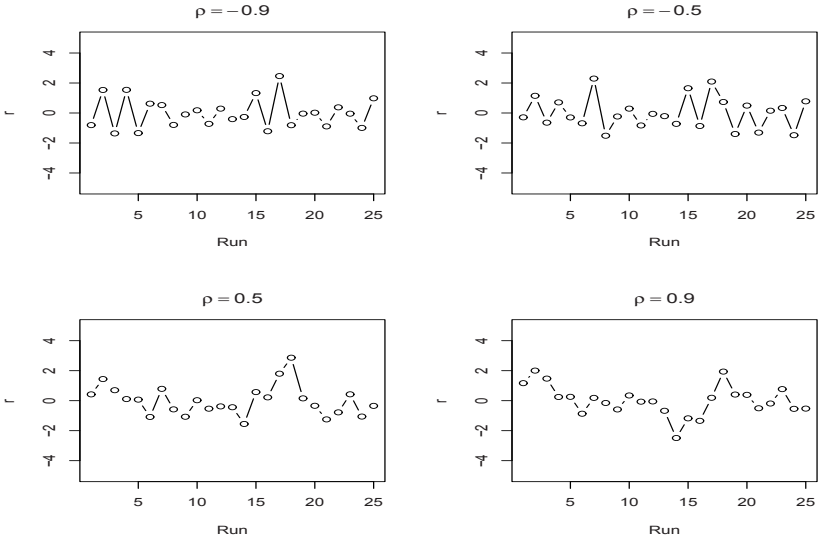

Fig. 13.9 Serial correlation standardized residual plots.

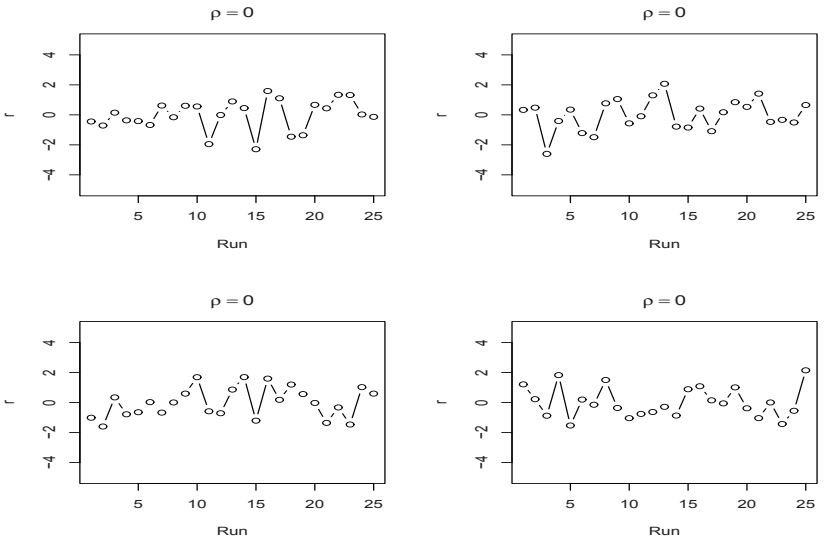

Fig. 13.10 Serial correlation standardized residual plots with uncorrelated data.

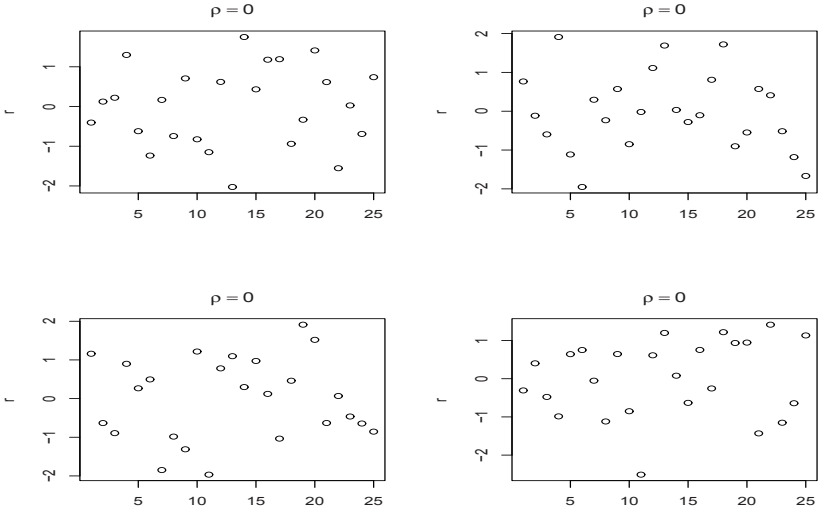

Fig. 13.11 Standardized residual plots when model assumptions are valid.

try to detect patterns in Figures 13.10 though 13.14 because the human eye is good at detecting/creating patterns, even though none exist in these plots.

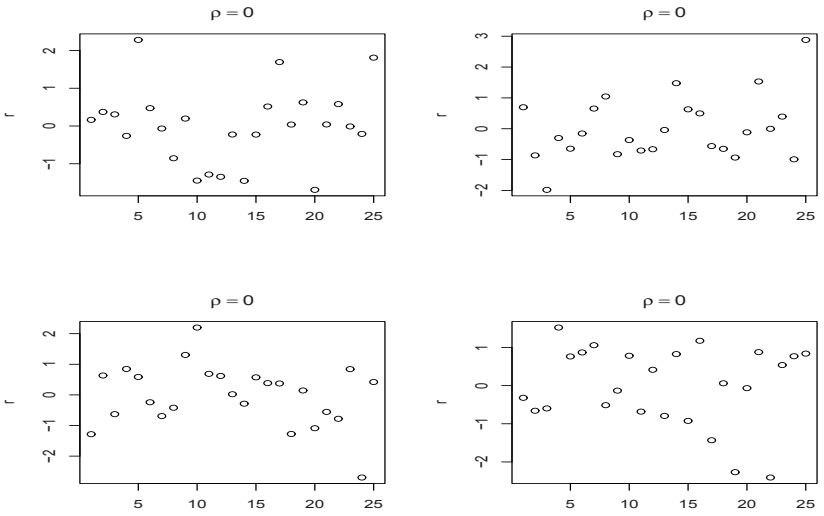

Fig. 13.12 Standardized residual plots when model assumptions are valid.

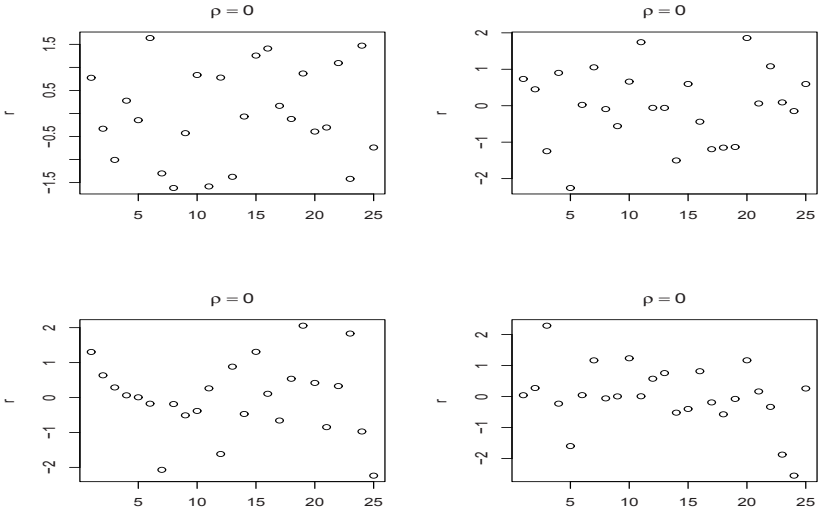

Fig. 13.13 Standardized residual plots when model assumptions are valid.

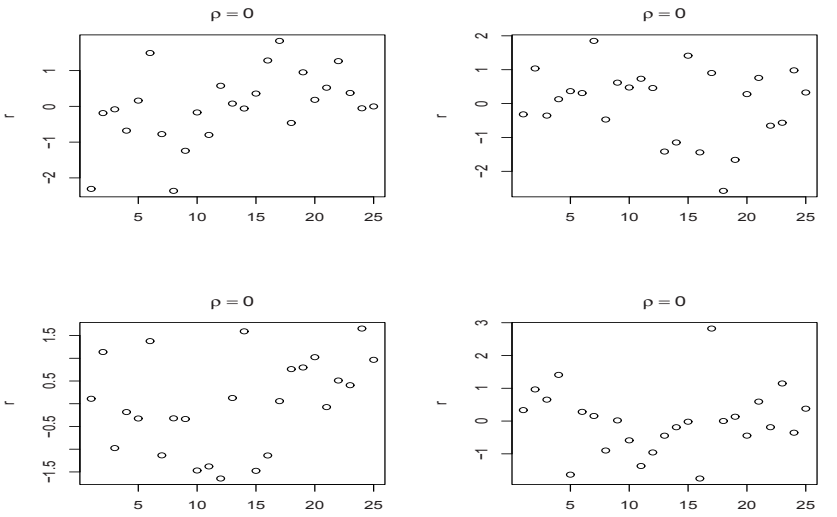

Fig. 13.14 Standardized residual plots when model assumptions are valid.

Durbin and Watson (1951) provided an approximate test for the hypothesis $\rho = 0$. The Durbin–Watson test statistic is

$$d = \sum_{i=1}^{n-1} (\hat{e}_{i+1} - \hat{e}_i)^2 \Big/ \sum_{i=1}^{n} \hat{e}_i^2.$$

Here d is an estimate of $\sum_{i=1}^{n-1}(e_{i+1} - e_i)^2 / \sum_{i=1}^{n} e_i^2$. For an AR(1) structure,

$$e_{i+1} - e_i = \rho e_i + v_{i+1} - e_i = (1-\rho)e_i + v_{i+1},$$

so we have

$$\mathrm{E}[e_{i+1} - e_i]^2 = \mathrm{E}[(1-\rho)e_i + v_{i+1}]^2 = (1-\rho)^2\sigma^2 + (1-\rho^2)\sigma^2 = 2(1-\rho)\sigma^2,$$

and

$$\mathrm{E}[e_i^2] = \sigma^2.$$

It follows that $\sum_{i=1}^{n-1}(\hat{e}_{i+1} - \hat{e}_i)^2$ should, for some constant K_1, estimate $K_1(1-\rho)\sigma^2$, and $\sum_{i=1}^{n} \hat{e}_i^2$ estimates $K_2\sigma^2$. d is a rough estimate of $(K_1/K_2)(1-\rho)$ or $K[1-\rho]$. If $\rho = 0$, d should be near K. If $\rho > 0$, d will tend to be small. If $\rho < 0$, d will tend to be large.

The exact distribution of d varies with X. For example, $\sum_{i=1}^{n-1}(\hat{e}_{i+1} - \hat{e}_i)^2$ is just a quadratic form in $\hat{e}$, say $\hat{e}A\hat{e} = Y'(I-M)A(I-M)Y$. It takes little effort to see that A is a very simple matrix. By Theorem 1.3.2,

$$\mathrm{E}(\hat{e}A\hat{e}) = \mathrm{tr}[(I-M)A(I-M)\mathrm{Cov}(Y)].$$

Thus, even the expected value of the numerator of d depends on the model matrix. Since the distribution depends on X, it is not surprising that the exact distribution of d varies with X.

Exercise 13.5 Show that d is approximately equal to $2(1-r_a)$, where r_a is the sample (auto)correlation between the pairs $(\hat{e}_{i+1}, \hat{e}_i)$ $i = 1, \ldots, n-1$.

13.4 Heteroscedasticity and Lack of Fit

Heteroscedasticity refers to having unequal variances. In particular, an independent heteroscedastic model has

$$\mathrm{Cov}(e) = \mathrm{Diag}(\sigma_i^2).$$

Lack of fit refers to having an incorrect model for $\mathrm{E}(Y)$. In Section 6.6 on testing lack of fit, we viewed this as having an insufficiently general model matrix. When lack of fit occurs, $\mathrm{E}(e) \equiv \mathrm{E}(Y - X\beta) \neq 0$. Both heteroscedasticity and lack of fit are diagnosed by plotting the standardized residuals against any variable of choice. The chosen variable may be case numbers, time sequence, any predictor variable included in the model, any predictor variable not included in the model, or the predicted values $\hat{Y} = MY$. If there is no lack of fit or heteroscedasticity, the residual plots should form a horizontal band. The plots in Section 3 with $\rho = 0.0$ are examples of such plots when the horizontal axis has equally spaced entries.

13.4.1 Heteroscedasticity

A horn-shaped pattern in a residual plot indicates that the variance of the observations is increasing or decreasing with the other variable.

EXAMPLE 13.4.1. Twenty-five i.i.d. $N(0,1)$ random variates, $z_1, \ldots, z_{25}$ were generated and y values were computed using (13.0.1) with $e_i = x_{i1}z_i/60$. The variance of the e_is increase as the x_{i1}s increase.

Figure 13.15 plots the standardized residuals $R1$ against $\hat{Y}$ and X_2. The plot of $R1$ versus $\hat{Y}$ shows something of a horn shape, but it opens to the left and is largely dependent on one large residual with a $\hat{y}$ of about 8.3. The plot against X_2 shows very little. It is difficult to detect any pattern in either plot. In (b) the two relatively small values of x_2 don't help. The top left component of Figure 13.16 plots R1 against X_1 where you can detect a pattern but, again, by no means an obvious one. The impression of a horn shape opening to the right is due almost entirely to one large residual near $x_1 = 70$. The remaining plots in Figure 13.16 as well as the plots

in Figures 13.17 and 13.18 are independent replications. Often you can detect the horn shape but sometimes you cannot.

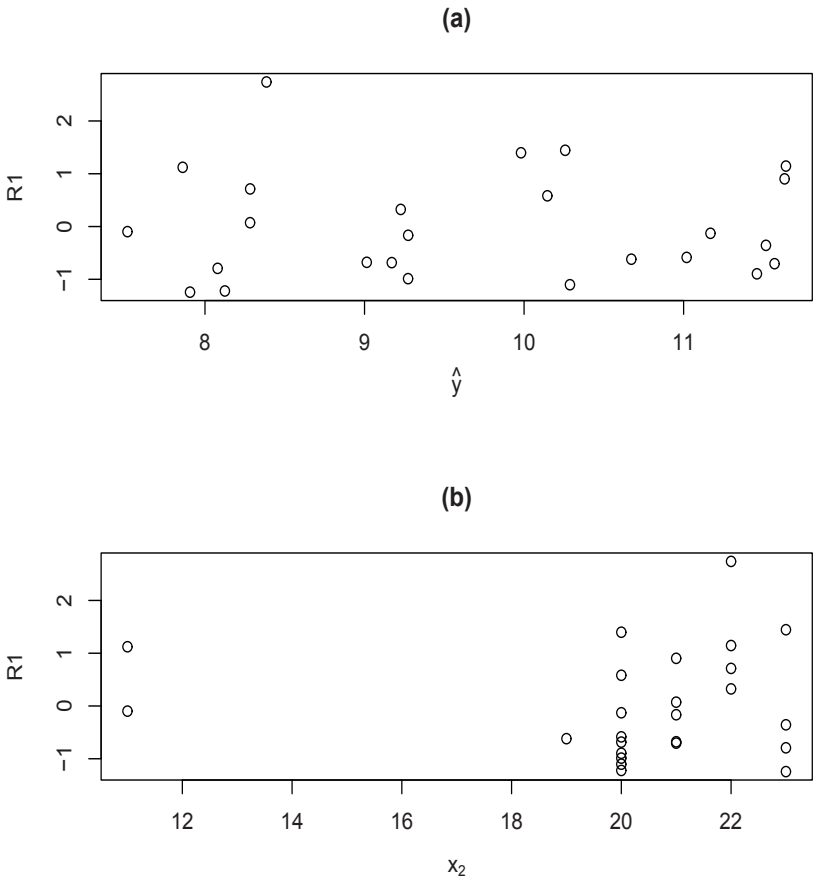

Fig. 13.15 Variance increasing with x_1.

EXAMPLE 13.4.2. To illustrate horn shapes that open to the left, Example 13.4.1 was repeated using $e_i = 60z_i/x_{i1}$. With these e_is, the variance decreases as x_1 increases. The plots are contained in Figures 13.19 through 13.21. In Figure 13.19(a) of $R1$ versus $\hat{Y}$, we see a horn opening to the right. Note that from (13.0.1), if x_2 is held constant, y increases as x_1 decreases. In the plot, x_2 is not being held constant, but the relationship still appears. There is little to see in Figure 13.19(b) $R1$ versus X_2. The plot of $R1$ versus X_1 in the top left of Figure 13.20 shows a horn opening to

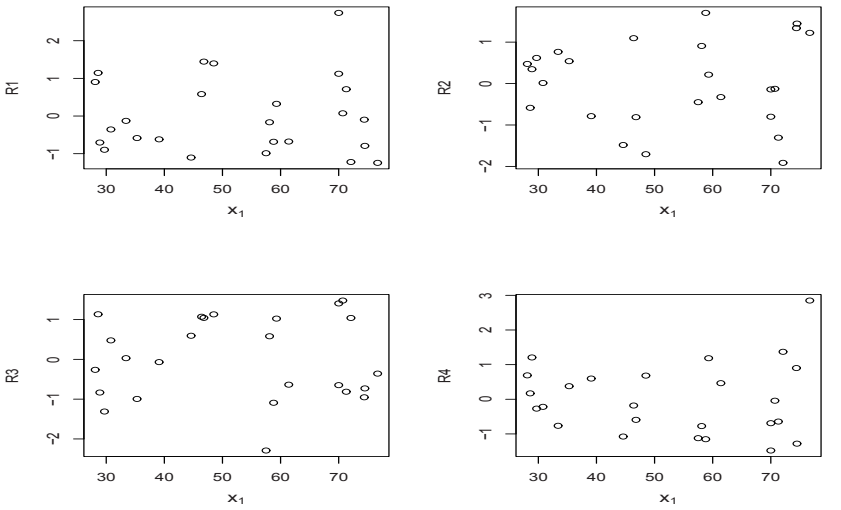

Fig. 13.16 Variance increasing with x_1.

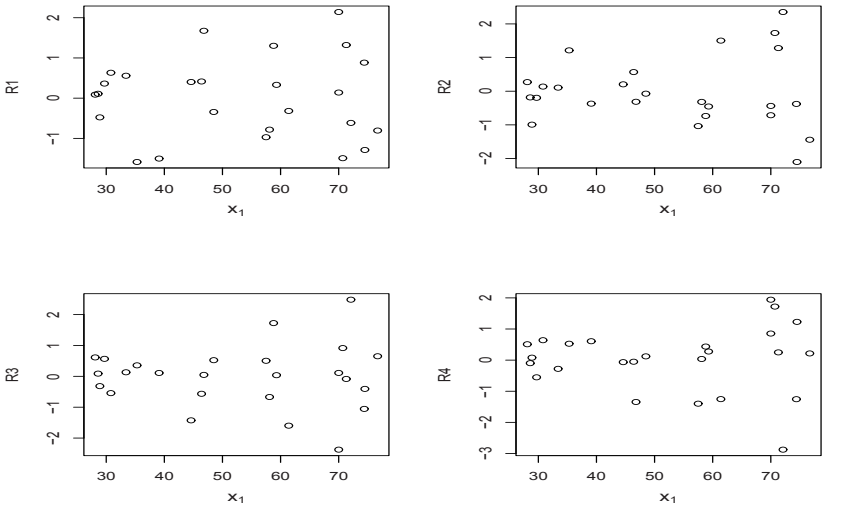

Fig. 13.17 Variance increasing with x_1.

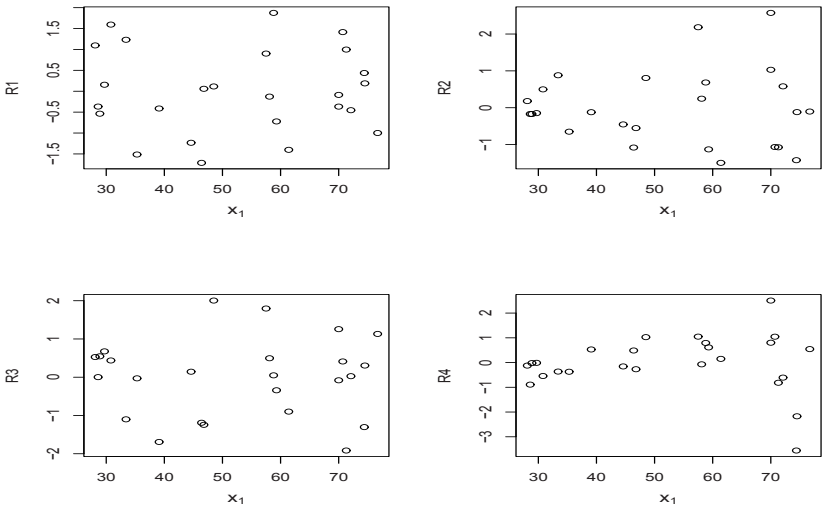

Fig. 13.18 Variance increasing with x_1.

the left. The remaining plots in Figure 13.20 as well as the plots in Figure 13.21 are independent replications.

Although plotting the residuals seems to be the standard method for examining heteroscedasticity of variances, Examples 13.4.1 and 13.4.2 indicate that residual plots are far from foolproof.

For one-way ANOVA models (and equivalent models such as two-way ANOVA with interaction), there are formal tests of heteroscedasticity available. The best known of these are Hartley's, Bartlett's, and Cochran's tests. The tests are based on the sample variances for the individual groups, say $s_1^2, \ldots, s_t^2$. Hartley's and Cochran's tests require equal sample sizes for each treatment group; Bartlett's test does not. Hartley's test statistic is $\max_i s_i^2 / \min_i s_i^2$. Cochran's test statistic is $\max_i s_i^2 / \sum_{i=1}^t s_i^2$. Bartlett's test statistic is $(n-t)\log \bar{s}_\cdot^2 - \sum_i (N_i - 1)\log s_i^2$. Descriptions of these and other tests can be found in Mandansky (1988).

All of these tests are based on the assumption that the data are normally distributed, and the tests are quite notoriously sensitive to the invalidity of that assumption. For nonnormal data, the tests frequently reject the hypothesis of all variances being equal, even when all variances are, in fact, equal. This is important because t and F tests tend not to be horribly sensitive to nonnormality. In other words, if the data are not normally distributed (and they never are), the data may be close enough to being normally distributed so that the t and F tests are approximately correct. However, the nonnormality may be enough to make Hartley's, Bartlett's,

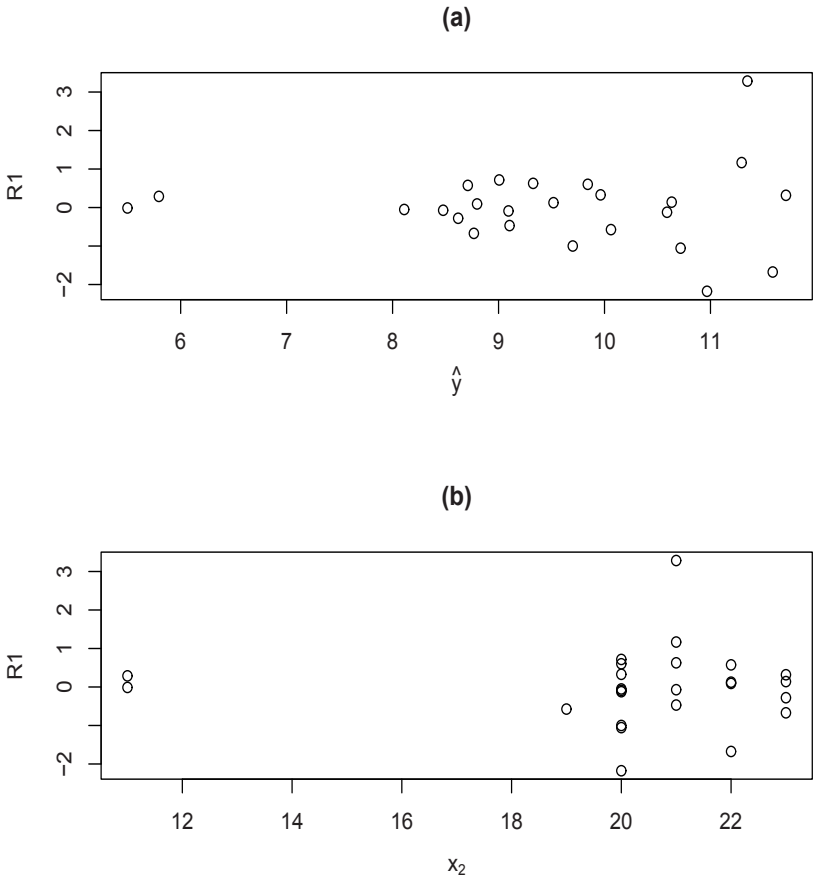

Fig. 13.19 Variance decreasing with x_1.

and Cochran's tests reject, so that the data analyst worries about a nonexistent problem of heteroscedasticity.

13.4.2 Lack of Fit

An additional use of residual plots is to identify lack of fit. The assumption is that $\mathrm{E}(e) = 0$, so any *systematic* pattern in the residuals (other than a horn shape) can indicate lack of fit. Most commonly, one looks for a linear or quadratic trend in the

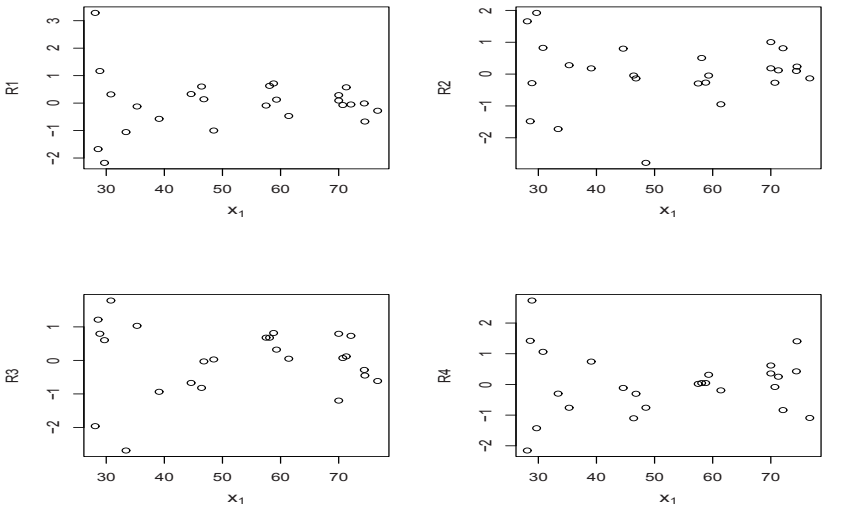

Fig. 13.20 Variance decreasing with x_1.

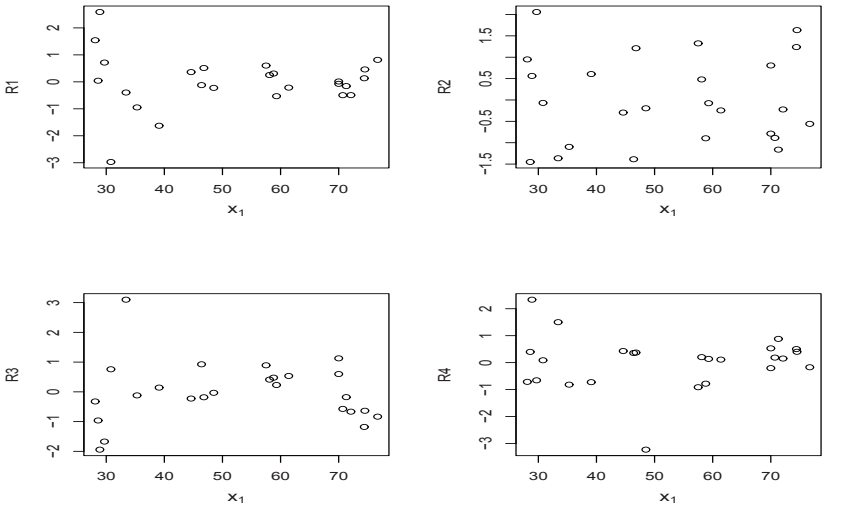

Fig. 13.21 Variance decreasing with x_1.

residuals. Such trends indicate the existence of effects that have not been removed from the residuals, i.e., effects that have not been accounted for in the model.

Theorems 6.3.3 and 6.3.6 indicate that the residuals should be uncorrelated with any function of the predictor variables when we have the best possible model. So a nonzero correlation between the residuals and any other variable, say z, indicates that something is wrong with the model. If z were not originally in the model, then it needs to be. A quadratic relationship between the residuals and z is indicative of a nonzero correlation between the residuals and z^2, although the linear relationship might be much clearer when plotting the residuals against a standardized version of z, say $z - \bar{z}$.

For examining heteroscedasticity, the standardized residuals need to be used because the ordinary residuals are themselves heteroscedastic. For examining lack of fit, the ordinary residuals are preferred but we often use the standardized residuals for convenience.

EXAMPLE 13.4.3. Data were generated using (13.0.1) and the incorrect model

$$y_i = \beta_0 + \beta_2 x_{i2} + e_i$$

was fitted. In the independently generated Figures 13.22 and 13.23, the ordinary residuals $\hat{e}$ and the standardized residuals R are plotted against x_1 in (a) and (b) to examine whether X_1 needs to be added to the model. The decreasing trends in the residual plots indicate that x_1 may be worth adding to the model.

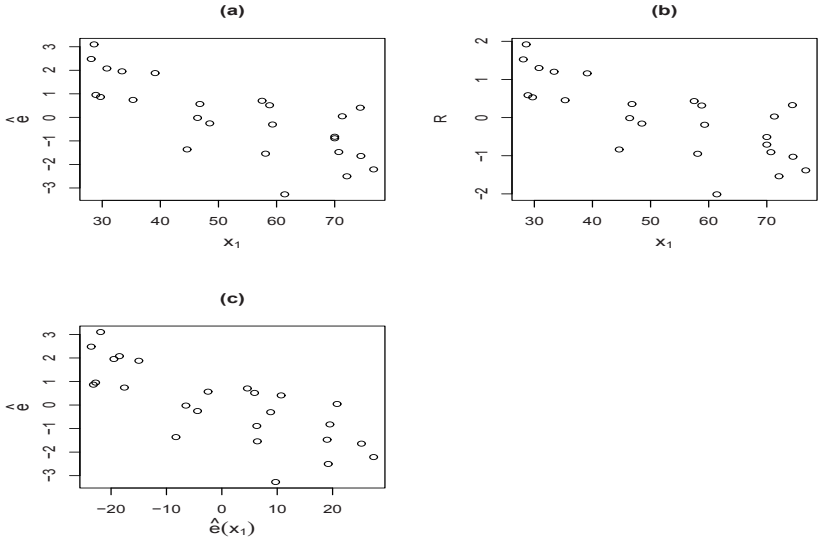

Fig. 13.22 Linear lack of fit plots.

Part (c) contains an *added variable plot*. To obtain it, find the ordinary residuals, say $\hat{e}(x_1)$, from fitting

$$x_{i1} = \gamma_0 + \gamma_2 x_{i2} + e_i.$$

By Exercise 9.2, a plot of $\hat{e}$ versus $\hat{e}(x_1)$ gives an exact graphical display of the effect of adding x_1 to the model.

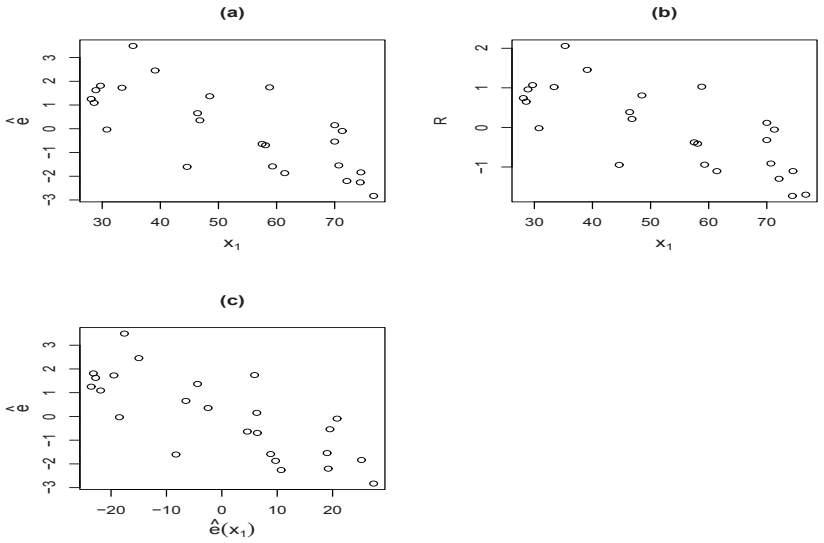

Fig. 13.23 Linear lack of fit plots.

The disadvantage of added variable plots is that it is time consuming to adjust the predictor variables under consideration for the variables already in the model. It is more convenient to plot residuals against predictor variables that have not been adjusted. As in Example 13.4.3, such plots are often informative but could, potentially, be misleading.

The final example in this section displays a quadratic lack of fit:

EXAMPLE 13.4.4. Data were generated by adding $0.005x_1^2$ to (13.0.1). The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

was fitted and standardized residuals R were obtained. Figures 13.24 and 13.25 are independent replications in which the standardized residuals are plotted against predicted values $\hat{y}$, against x_2, and against x_1. The quadratic trend appears clearly in the plots of residuals versus $\hat{y}$ and x_1, and even seems to be hinted at in the plot versus x_2.

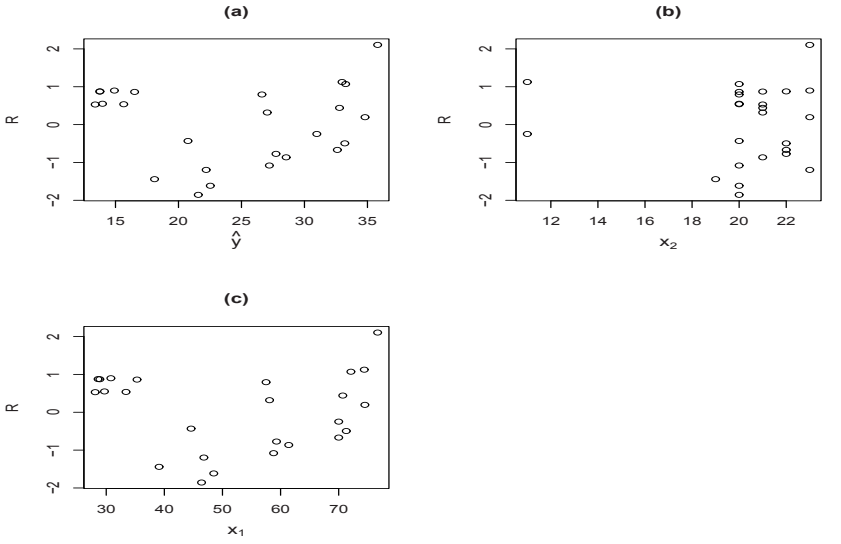

Fig. 13.24 Quadratic lack of fit plots.

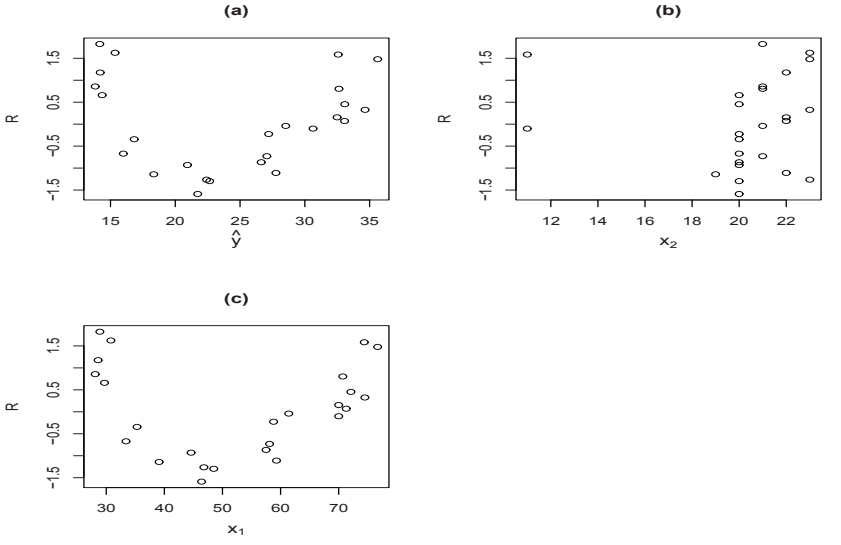

Fig. 13.25 Quadratic lack of fit plots.

In Figures 13.24 and 13.25, it would not be possible to have a linear effect. The residuals are being plotted against variables x_1 and x_2 that are already included in the fitted model as well as $\hat{y}$ which is a linear combination of x_1 and x_2. Any linear effect of these variables would be eliminated by the fitting procedure. However, plotting the residuals against a variable *not* included in the model does allow the possibility of seeing a linear effect, e.g., plotting against $(x_{i1} - \bar{x}_{\cdot 1})^2$.

EXAMPLE 13.4.5. As mentioned earlier, lack of independence and lack of fit are closely related. In Figures 13.7 through 13.9, we plotted residuals against case numbers. When observations are taken sequentially, the case numbers can be thought of as a measure of time. There was *no* lack of fit present in the fitted models for these plots. Nonetheless, the serial correlation can cause the plots to look like the models are lacking linear or quadratic effects in time, respectively, especially when $\rho = 0.9$.

Tests for lack of fit were discussed in Section 6.6. Although no examples of it have been presented, it should be noted that it is possible to have both lack of fit and heteroscedasticity in the same plot.

13.5 Updating Formulae and Predicted Residuals

Frequently, it is of interest to see how well the data could predict y_i if the ith case were left out when the model was fitted. The difference between y_i and the estimate $\hat{y}_{[i]}$ with the ith case deleted is called either the *predicted residual* or the *deleted residual*. The computations for fitting the model with the ith case left out can be performed by using simple updating formulae on the complete model. We give several of these formulae.

Let $X_{[i]}$ and $Y_{[i]}$ be X and Y with the ith row deleted. Write x_i' for the ith row of X and

$$\hat{\beta}_{[i]} = (X_{[i]}'X_{[i]})^{-1}X_{[i]}'Y_{[i]}$$

for the estimate of β without the ith case. The predicted residual is defined as

$$\hat{e}_{[i]} = y_i - x_i'\hat{\beta}_{[i]}.$$

The predicted residuals are useful in checking for outliers. They are also used for model selection. The Predicted REsidual Sum of Squares (PRESS) is defined as

$$\text{PRESS} = \sum_{i=1}^{n} \hat{e}_{[i]}^2.$$

Models with relatively low values of the PRESS statistic should be better than models with high PRESS statistics. It is tempting to think that PRESS is a more valid measure of how well a model fits than SSE, because PRESS predicts values not used in fitting the model. This reasoning may seem less compelling after the updating formula for the predicted residuals has been established.

The predicted residuals can also be used to check normality, heteroscedasticity, and lack of fit in the same way that the usual residuals are used. For these purposes they should be standardized. Their variances are

$$\begin{aligned}\operatorname{Var}(\hat{e}_{[i]}) &= \sigma^2 + \sigma^2 x_i'(X_{[i]}'X_{[i]})^{-1}x_i \\ &= \sigma^2\left[1 + x_i'(X_{[i]}'X_{[i]})^{-1}x_i\right].\end{aligned}$$

A reasonable estimate of σ^2 is $MSE_{[i]}$, the mean squared error for the model with the ith case deleted. Alternatively, σ^2 could be estimated with the regular MSE. If MSE is used, then the standardized predicted residuals are identical to the standardized residuals (see Exercise 13.6). Standardized predicted residuals will be discussed again in Section 6. A more useful formula for $\operatorname{Var}(\hat{e}_{[i]})$ is given in Proposition 13.5.4.

We now present a series of results that establish the updating formulae for models with one deleted case.

Proposition 13.5.1. Let A be a $p \times p$ nonsingular matrix, and let a and b be $q \times p$ rank q matrices. Then, if all inverses exist,

$$(A + a'b)^{-1} = A^{-1} - A^{-1}a'(I + bA^{-1}a')^{-1}bA^{-1}.$$

PROOF. This is a special case of Theorem B.56. □

The application of Proposition 13.5.1 is

Corollary 13.5.2.

$$(X_{[i]}'X_{[i]})^{-1} = (X'X)^{-1} + [(X'X)^{-1}x_i x_i'(X'X)^{-1}]/[1 - x_i'(X'X)^{-1}x_i].$$

PROOF. The corollary follows from noticing that $X_{[i]}'X_{[i]} = (X'X - x_i x_i')$. □

Proposition 13.5.3. $\hat{\beta}_{[i]} = \hat{\beta} - [(X'X)^{-1}x_i\hat{e}_i]/(1 - m_{ii})$.

PROOF. First, note that $x_i'(X'X)^{-1}x_i = m_{ii}$ and $X_{[i]}'Y_{[i]} = X'Y - x_i y_i$. Now, from Corollary 13.5.2,

$$\begin{aligned}\hat{\beta}_{[i]} &= (X_{[i]}'X_{[i]})^{-1}X_{[i]}'Y_{[i]} \\ &= (X_{[i]}'X_{[i]})^{-1}(X'Y - x_i y_i) \\ &= \hat{\beta} - (X'X)^{-1}x_i y_i + \left[(X'X)^{-1}x_i x_i'\hat{\beta} - (X'X)^{-1}x_i x_i'(X'X)^{-1}x_i y_i\right]/(1 - m_{ii}).\end{aligned}$$

Writing $(X'X)^{-1}x_i y_i$ as $(X'X)^{-1}x_i y_i/(1 - m_{ii}) - m_{ii}(X'X)^{-1}x_i y_i/(1 - m_{ii})$, it is easily seen that

$$\begin{aligned}\hat{\beta}_{[i]} &= \hat{\beta} - [(X'X)^{-1}x_i(y_i - x_i'\hat{\beta})]/(1-m_{ii}) \\ &\quad + \left[m_{ii}(X'X)^{-1}x_iy_i - (X'X)^{-1}x_im_{ii}y_i\right]/(1-m_{ii}) \\ &= \hat{\beta} - [(X'X)^{-1}x_i\hat{e}_i]/(1-m_{ii}).\end{aligned}$$

□

The predicted residuals can now be written in a simple way.

Proposition 13.5.4.

(a) $\hat{e}_{[i]} = \hat{e}_i/(1-m_{ii})$.
(b) $\mathrm{Var}(\hat{e}_{[i]}) = \sigma^2/(1-m_{ii})$.

PROOF.

(a)
$$\begin{aligned}\hat{e}_{[i]} &= y_i - x_i'\hat{\beta}_{[i]} \\ &= y_i - x_i'\left[\hat{\beta} - \frac{(X'X)^{-1}x_i\hat{e}_i}{1-m_{ii}}\right] \\ &= \hat{e}_i + m_{ii}\hat{e}_i/(1-m_{ii}) \\ &= \hat{e}_i/(1-m_{ii}).\end{aligned}$$

(b) This follows from having $\hat{e}_{[i]} = \hat{e}_i/(1-m_{ii})$ and $\mathrm{Var}(\hat{e}_i) = \sigma^2(1-m_{ii})$. □

The PRESS statistic can now be written as

$$\mathrm{PRESS} = \sum_{i=1}^{n} \hat{e}_i^2/(1-m_{ii})^2.$$

The value of $\hat{e}_i^2/(1-m_{ii})^2$ will usually be large when m_{ii} is near 1. Model selection with PRESS puts a premium on having models in which observations with extremely high leverage are fitted very well. As will be discussed in Chapter 14, when fitting a model after going through a procedure to select a good model, the fitted model tends to be very optimistic in the sense of indicating much less variability than is appropriate. Model selection using the PRESS statistic tends to continue that phenomenon, cf. Picard and Cook (1984) and Picard and Berk (1990).

Later, we will also need the sum of squares for error with the ith case deleted, say $SSE_{[i]}$.

Proposition 13.5.5. $SSE_{[i]} = SSE - \hat{e}_i^2/(1-m_{ii})$.

PROOF. By definition,

$$\begin{aligned}SSE_{[i]} &= Y_{[i]}'Y_{[i]} - Y_{[i]}'X_{[i]}(X_{[i]}'X_{[i]})^{-1}X_{[i]}'Y_{[i]} \\ &= (Y'Y - y_i^2) - Y_{[i]}'X_{[i]}\hat{\beta}_{[i]}.\end{aligned}$$

The second term can be written

$$\begin{aligned} Y'_{[i]}X_{[i]}\hat{\beta}_{[i]} &= (Y'X - y_i x_i')\left\{\hat{\beta} - [(X'X)^{-1}x_i\hat{e}_i]/(1-m_{ii})\right\} \\ &= Y'X\hat{\beta} - y_i x_i'\hat{\beta} - x_i'\hat{\beta}\hat{e}_i/(1-m_{ii}) + y_i m_{ii}\hat{e}_i/(1-m_{ii}) \\ &= Y'MY - y_i x_i'\hat{\beta} + y_i\hat{e}_i/(1-m_{ii}) - x_i'\hat{\beta}\hat{e}_i/(1-m_{ii}) \\ &\quad - y_i\hat{e}_i/(1-m_{ii}) + y_i m_{ii}\hat{e}_i/(1-m_{ii}) \\ &= Y'MY - y_i x_i'\hat{\beta} + \hat{e}_i^2/(1-m_{ii}) - y_i\hat{e}_i \\ &= Y'MY + \hat{e}_i^2/(1-m_{ii}) - y_i^2. \end{aligned}$$

Therefore,

$$\begin{aligned} SSE_{[i]} &= Y'Y - y_i^2 - [Y'MY + \hat{e}_i^2/(1-m_{ii}) - y_i^2] \\ &= Y'(I-M)Y - \hat{e}_i^2/(1-m_{ii}) \\ &= SSE - \hat{e}_i^2/(1-m_{ii}). \end{aligned}$$

□

Exercise 13.6 Show that the standardized predicted residuals with σ^2 estimated by MSE are the same as the standardized residuals.

13.6 Outliers and Influential Observations

Realistically, the purpose of fitting a linear model is to get a (relatively) succinct summary of the important features of the data. Rarely is the chosen linear model really correct. Usually, the linear model is no more than a rough approximation to reality.

Outliers are cases that do not seem to fit the chosen linear model. There are two kinds of outliers. Outliers may occur because the predictor variables for the case are unlike the predictor variables for the other cases. These are cases with high leverage. If we think of the linear model as being an approximation to reality, the approximation may be quite good within a certain range of the predictor variables, but poor outside that range. A few cases that fall outside the range of good approximation can greatly distort the fitted model and lead to a bad fit, even on the range of good approximation. Outliers of this kind are referred to as *outliers in the design space (estimation space).*

The other kind of outliers are those due to bizarre values of the dependent variable. These may occur because of gross measurement error, or from recording the data incorrectly. Not infrequently, data are generated from a mixture process. In other words, the data fit one pattern most of the time, but occasionally data with a different pattern are generated. Often it is appropriate to identify the different kinds of data and model them separately. If the vast majority of data fit a common pattern, there may not be enough of the rare observations for a complete analysis; but it is

still important to identify such observations. In fact, these rare observations can be more important than all of the other data.

Not only is it important to be able to identify outliers, but it must also be decided whether such observations should be included when fitting the model. If they are left out, one gets an approximation to what usually happens, and one must be aware that something unusual will happen every so often.

Outliers in the design space are identified by their leverages, the m_{ii}s. Bizarre values of y_i can often be identified by their standardized residuals. Large standardized residuals indicate outliers. Typically, these are easily spotted in residual plots. However, if a case with an unusual y_i also has high leverage, the standardized residual may not be large. If there is only one bizarre value of y_i, it should be easy to identify by examining all the cases with either a large standardized residual or high leverage. With more than one bizarre value, they *may* mask each other. (I believe that careful examination of the leverages will almost always identify possible masking, cf. the comments in Section 1 on gaps in the leverage values.)

An alternative to examining the standardized residuals is to examine the *standardized predicted residuals*, also known as the *standardized deleted residuals*, the *t residuals*, and sometimes (as in the R programming language) the *Studentized residuals*. The standardized predicted residuals are

$$t_i = \frac{\hat{e}_{[i]}}{\sqrt{MSE_{[i]}/(1-m_{ii})}} = \frac{y_i - x_i'\hat{\beta}_{[i]}}{\sqrt{MSE_{[i]}/(1-m_{ii})}}.$$

Since y_i, $\hat{\beta}_{[i]}$, and $MSE_{[i]}$ are independent,

$$t_i \sim t(n-p-1),$$

where $p = r(X)$. This allows a formal t test for whether the value y_i is consistent with the rest of the data. Actually, this procedure is equivalent to examining the standardized residuals, but using the $t(n-p-1)$ distribution is more convenient than using the appropriate distribution for the r_is, cf. Cook and Weisberg (1982).

When all the values t_i, $i = 1,\ldots,n$, are computed, the large values will naturally be singled out for testing. The appropriate test statistic is actually $\max_i |t_i|$. The null distribution of this statistic is quite different from a $t(n-p-1)$. Fortunately, Bonferroni's inequality provides an appropriate, actually a conservative, P value by multiplying the P value from a $t(n-p-1)$ distribution by n. Alternatively, for an α level test, use a critical value of $t(1-\alpha/2n, n-p-1)$.

If y_i corresponds to a case with extremely high leverage, the standard error for the predicted residual, $\sqrt{MSE_{[i]}/(1-m_{ii})}$, will be large, and it will be difficult to reject the t test. Recall from Example 13.1.6 that under condition (3) the y value for case six is clearly discordant. Although the absolute t value is quite large, $t_6 = -5.86$ with $m_{66} = 0.936$, it is smaller than one might expect, considering the obvious discordance of case six. In particular, the absolute t value is smaller than the critical point for the Bonferroni method with $\alpha = 0.05$. (The critical point is

$t(1-0.025/6,3) = 6.23$.) Of course, with three degrees of freedom, the power of this test is very small. A larger α level would probably be more appropriate. Using the Bonferroni method with $\alpha = 0.10$ leads to rejection.

The updating formula for t_i is

Proposition 13.6.1.

$$t_i = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}.$$

PROOF. Using the updating formulae of Section 5,

$$\begin{aligned}
t_i &= \hat{e}_{[i]} \Big/ \sqrt{MSE_{[i]}/(1-m_{ii})} \\
&= \hat{e}_{[i]} \sqrt{(1-m_{ii})} \Big/ \sqrt{MSE_{[i]}} \\
&= r_i \sqrt{MSE} \Big/ \sqrt{MSE_{[i]}} \\
&= r_i \sqrt{(n-p-1)/(n-p)} \sqrt{SSE/SSE_{[i]}} \\
&= r_i \sqrt{(n-p-1)/(n-p)} \sqrt{SSE/[SSE - \hat{e}_i^2/(1-m_{ii})]} \\
&= r_i \sqrt{(n-p-1)/(n-p)} \sqrt{1/[1-r_i^2/(n-p)]} \\
&= r_i \sqrt{(n-p-1)/(n-p-r_i^2)}.
\end{aligned}$$

□

As indicated earlier, t_i really contains the same information as r_i.

A test that a given set of y values does not fit the model is easily available from general linear model theory. Suppose that the r observations $i = n-r+1, \ldots, n$ are suspected of being outliers. The model $Y = X\beta + e$ can be written with

$$Y = \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}, \quad X = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}, \quad e = \begin{bmatrix} e_0 \\ e_1 \end{bmatrix},$$

where Y_1, X_1, and e_1 each have r rows. If $Z = \begin{bmatrix} 0 \\ I_r \end{bmatrix}$, then the model with the possible outliers deleted

$$Y_0 = X_0\beta + e_0 \tag{1}$$

and the model

$$Y = X\beta + Z\gamma + e \tag{2}$$

are equivalent for estimating β and σ^2. A test of the reduced model $Y = X\beta + e$ against the full model $Y = X\beta + Z\gamma + e$ is rejected if

$$\frac{(SSE - SSE_0)/r}{MSE_0} > F(1-\alpha, r, n-p-r).$$

If the test is rejected, the r observations appear to contain outliers. Note that this procedure is essentially the same as Utts's Rainbow Test for lack of fit discussed in Section 6.6. The difference is in how one identifies the cases to be eliminated.

In my opinion, the two most valuable tools for identifying outliers are the m_{ii}s and the t_is. It would be unusual to have outliers in the design space without large values of the m_{ii}s. Such outliers would have to be "far" from the other data without the Mahalanobis distance being large. For bizarre values of the y_is, it is useful to determine whether the value is so bizarre that it could not reasonably come from the model under consideration. The t_is provide a test of exactly what is needed.

Cook (1977) presented a distance measure that combines the standardized residual with the leverage to get a single measure of the influence a case has on the fit of the regression model. Cook's distance (C_i) measures the statistical distance between $\hat{\beta}$ and $\hat{\beta}_{[i]}$. It is defined as

$$C_i = \frac{(\hat{\beta}_{[i]} - \hat{\beta})'(X'X)(\hat{\beta}_{[i]} - \hat{\beta})}{pMSE}.$$

Written as a function of the standardized residual and the leverage, Cook's distance is:

Proposition 13.6.2. $C_i = r_i^2[m_{ii}/p(1-m_{ii})]$.

Exercise 13.7 Prove Proposition 13.6.2.

From Proposition 13.6.2 it can be seen that Cook's distance takes the size of the standardized residual and rescales it based on the leverage of the case. For an extremely high leverage case, the squared standardized residual gets multiplied by a very large number. For low leverage cases the multiplier is very small. Another interpretation of Cook's distance is that it is a standardized version of how far the predicted values $X\hat{\beta}$ move when the ith case is deleted.

In Section 1, after establishing that the m_{ii}s were a reasonable measure of leverage, it was necessary to find guidelines for what particular values of m_{ii} meant. This can also be done for Cook's distance. Cook's distance can be calibrated in terms of confidence regions. Recall that a $(1-\alpha)100\%$ confidence region for β is

$$\left\{ \beta \middle| \frac{(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})}{pMSE} < F(1-\alpha, p, n-p) \right\}.$$

If $C_i \doteq F(0.75, p, n-p)$, then deleting the ith case moves the estimate of β to the edge of a 75% confidence region for β based on $\hat{\beta}$. This is a very substantial move. Since $F(0.5, p, n-p) \doteq 1$, any case with $C_i > 1$ probably has above average influence. Note that C_i does not actually have an F distribution. While many people consider such calibrations a necessity, other people, including the author, prefer sim-

ply to examine those cases with distances that are substantially larger than the other distances.

Cook's distance can be modified in an obvious way to measure the influence of any set of observations. Cook and Weisberg (1982) give a more detailed discussion of all of the topics in this section.

Updating formulae and case deletion diagnostics for linear models with general covariance matrices are discussed by Christensen, Johnson, and Pearson (1992, 1993), Christensen, Pearson, and Johnson (1992), and by Haslett and Hayes (1998) and Martin (1992). A nice review of these procedures is given by Shi and Chen (2009).

Haslett (1999) establishes two results that can greatly simplify case deletion diagnostics for correlated data. First, he shows that an analysis based on $Y_{[i]}$, $X_{[i]}$, and $V_{[ii]}$ (the covariance matrix with the ith row and column removed) is the same as an analysis based on $\tilde{Y}(i)$, X, and V where

$$\tilde{Y}(i) = \begin{bmatrix} \hat{y}_i(Y_{[i]}) \\ Y_{[i]} \end{bmatrix}$$

and $\hat{y}_i(Y_{[i]})$ is the BLUP of y_i based on $Y_{[i]}$. Second, he shows that there is a relatively simple way to find a matrix P_i such that $\tilde{Y}(i) = P_i Y$.

13.7 Transformations

If the residuals suggest nonnormality, heteroscedasticity of variances, or lack of fit, a transformation of the y_is may eliminate the problem. Cook and Weisberg (1982) and Atkinson (1985) give extensive discussions of transformations. Only a brief review is presented here.

Picking a transformation is often a matter of trial and error. Different transformations are tried until one is found for which the residuals seem reasonable. Three more systematic methods of choosing transformations will be discussed: Box–Cox power transformations, variance stabilizing transformations, and the generalized least squares approach of Grizzle, Starmer, and Koch (1969).

Box and Cox (1964) suggested a systematic method of picking transformations. They suggested using the family of transformations

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

and choosing λ by maximum likelihood. Convenient methods of executing this procedure are discussed in Cook and Weisberg (1982), Weisberg (1985), and Christensen (1996a).

If the distribution of the y_is is known, the commonly used *variance stabilizing transformations* can be tried (cf. Rao, 1973, Section 6g, or Christensen, 1996a). For example,

if $y_i \sim \text{Binomial}(N_i, p_i)$,	use $\arcsin(\sqrt{y_i/N_i})$,
if $y_i \sim \text{Poisson}(\lambda_i)$,	use $\sqrt{y_i}$,
if y_i has $\sigma_i/\text{E}(y_i)$ constant,	use $\log(y_i)$.

More generally, $\arcsin(\sqrt{y_i/N_i})$ can be tried for any problem where y_i is a count between 0 and N_i or a proportion, $\sqrt{y_i}$ can be used for any problem where y_i is a count, and $\log(y_i)$ can be used if y_i is a count or amount.

The transformation $\log(y_i)$ is also frequently used because, for a linear model in $\log(y_i)$, the additive effects of the predictor variables transform to multiplicative effects on the original scale. If multiplicative effects seem reasonable, the log transformation may be appropriate.

As an alternative to the variance stabilizing transformations, there exist generalized linear models specifically designed for treating binomial and Poisson data. For Poisson data there exists a well developed theory for fitting log-linear models. One branch of the theory of log-linear models is the theory of logistic regression, which is used to analyze binomial data. As shown by Grizzle, Starmer, and Koch (1969), generalized least squares methods can be used to fit log-linear models to Poisson data and logistic regression models to binomial data. The method involves both a transformation of the dependent variable and weights. The appropriate transformation and weights are:

Distribution of y_i	Transformation	Weights
$\text{Poisson}(\lambda_i)$	$\log y_i$	y_i
$\text{Binomial}(N_i, p_i)$	$\log(y_i/[N_i - y_i])$	$y_i(N_i - y_i)/N_i$

With these weights, the asymptotic variance of the transformed data is 1.0. Standard errors for regression coefficients are computed as usual except that no estimate of σ^2 is required (σ^2 is known to be 1). Since σ^2 is known, t tests and F tests are replaced with normal tests and chi-squared tests. In particular, if the linear model fits the data and the observations y_i are large, the *SSE* has an asymptotic chi-squared distribution with the usual degrees of freedom. If the *SSE* is too large, a lack of fit is indicated. Tests of various models can be performed by comparing the difference in the *SSE*s for the model. The difference in the *SSE*s has an asymptotic chi-squared distribution with the usual degrees of freedom. If the difference is too large, then the smaller model is deemed inadequate. Unlike the lack of fit test, these model comparison tests are typically valid if n is large even when the individual y_is are not. Fitting log-linear and logistic models both by generalized least squares and by maximum likelihood is discussed in detail by Christensen (1997).

Exercise 13.8 The data given below were first presented by Brownlee (1965) and have subsequently appeared in Daniel and Wood (1980), Draper and Smith (1998), and Andrews (1974), among other places. The data consist of measurements

taken on 21 successive days at a plant that oxidizes ammonia into nitric acid. The dependent variable y is stack loss. It is 10 times the percentage of ammonia that is lost (in the form of unabsorbed nitric oxides) during the oxidation process. There are three predictor variables: x_1, x_2, and x_3. The first predictor variable is air flow into the plant. The second predictor variable is the cooling water temperature as it enters the countercurrent nitric oxide absorption tower. The third predictor variable is a coded version of the nitric acid concentration in the absorbing liquid. Analyze these data giving special emphasis to residual analysis and influential observations.

Obs.	x_1	x_2	x_3	y	Obs.	x_1	x_2	x_3	y
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

Exercise 13.9 For testing whether one observation y_i is an outlier, show that the F statistic is equal to the squared standardized predicted residual.