CHAPTER 2

Checking the Straight Line Fit

We discuss basic methods of checking a fitted regression model. Although we talk about these in terms of fitting a straight line, the basic methods apply generally whenever a linear model is fitted, no matter how many predictors there are. Other techniques too advanced for our current context are given in Chapter 8. Here we examine the following:

1. The lack of fit F-test when the data contain *repeat observations*, that is, when *pure error* is available (Sections 2.1 and 2.2).
2. Basic visual checks that can be made on the residuals $e_i = Y_i - \hat{Y}_i$ (Sections 2.3–2.6).
3. The Durbin–Watson test for checking serial correlation (Section 2.7).

2.1. LACK OF FIT AND PURE ERROR

General Discussion of Variance and Bias

We have already remarked that the fitted regression line is a calculated line based on a certain model or assumption, an assumption we should not blindly accept but should *tentatively entertain*. In certain circumstances we can check whether or not the model is correct. First, we can examine the consequences of an incorrect model. Let us recall that $e_i = Y_i - \hat{Y}_i$ is the *residual* at X_i. This is the amount by which the actual observed value Y_i differs from the fitted value $\hat{Y}_i$. As shown in Section 1.2, $\Sigma e_i = 0$. The residuals contain all available information on the way in which the model fitted fails to properly explain the observed variation in the dependent variable Y. Let $\eta_i = E(Y_i)$ denote the value given by the true model, whatever it is, at $X = X_i$. Then we can write

$$\begin{aligned} Y_i - \hat{Y}_i &= (Y_i - \hat{Y}_i) - E(Y_i - \hat{Y}_i) + E(Y_i - \hat{Y}_i) \\ &= \{(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i))\} + (\eta_i - E(\hat{Y}_i)) \\ &= q_i + B_i, \end{aligned}$$

say, where

$$q_i = \{(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i))\}, \qquad B_i = \eta_i - E(\hat{Y}_i).$$

The quantity B_i is the bias error at $X = X_i$. If the model is correct, then $E(\hat{Y}_i) = \eta_i$ and B_i is zero. If the model is not correct, $E(\hat{Y}_i) \neq \eta_i$ and B_i is not zero but has a value that depends on the true model and the value of X_i. The quantity q_i is a ran-

dom variable that has zero mean since $E(q_i) = E(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) = \eta_i - E(\hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) = 0$, and this is true whether the model is correct or not, that is, whether $E(\hat{Y}_i) = \eta_i$ or not.

The q_i, it can be shown, are correlated, and the quantity $q_1^2 + q_2^2 + \cdots + q_n^2$ has expected or mean value $(n - 2)\sigma^2$, where $V(Y_i) = V(\epsilon_i) = \sigma^2$ is the error variance. From this it can be shown further that the residual mean square value

$$\frac{1}{n-2}\left\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right\} \tag{2.1.1}$$

has expected or mean value σ^2 if the postulated model is of the correct form, or $\sigma^2 + \Sigma B_i^2/(n - 2)$ if the model is not correct. If the model is correct, that is, if $B_i = 0$, then the residuals are (correlated) random deviations q_i and the residual mean square can be used as an estimate of the error variance σ^2.

However, if the model is not correct, that is, if $B_i \neq 0$, then the residuals contain both random (q_i) and systematic (B_i) components. We can refer to these as the variance error and bias error components of the residuals, respectively. Also, the residual mean square will tend to be inflated and will no longer provide a satisfactory measure of the random variation present in the observations. (Since, however, the mean square is a random variable it may, by chance, not have a large value even when bias does exist. For some similar work on the general regression case see Section 10.2.)

How Big is σ^2?

In the simple case of fitting a straight line, bias error can usually be detected merely by examining a plot of the data. When the model is more complicated and/or involves more variables this may not be possible. If a prior estimate of σ^2 is available (by "prior estimate" we mean one obtained from previous experience of the variation in the situation being studied) we can see (or test by an F-test) whether or not the residual mean square is significantly greater than this prior estimate. If it is significantly greater we say that there is lack of fit and we would reconsider the model, which would be inadequate in its present form. If no prior estimate of σ^2 is available, but repeat measurements of Y (i.e., two or more measurements) have been made at the same value of X, we can use these repeats to obtain an estimate of σ^2. Such an estimate is said to represent "pure error" because, if the setting of X is identical for two observations, only the random variation can influence the results and provide differences between them. Such differences will usually then provide an estimate of σ^2 which is much more reliable than we can obtain from any other source. For this reason, it is sensible when designing experiments to arrange for repeat observations.

Genuine Repeats Are Needed

It is important to understand that repeated runs must be genuine repeats and not just repetitions of the same reading. For example, suppose we were attempting to relate, by regression methods, Y = intelligence quotient to X = height of person. A genuine repeat point would be obtained if we measured the separate IQs of two people of exactly the same height. If, however, we measure the IQ of one person of some specified height *twice*, this would not be a genuine repeat point in our context but merely a "reconfirmed" single point. The latter would, it is true, supply information on the variation of the testing method, which is part of the variation σ^2, but it would *not* provide information on the variation in IQ between people of the same height,

which is the σ^2 of our problem. In chemical experiments, a succession of readings made during steady-state running does not provide genuine repeat points. However, if a certain set of conditions was reset anew, after intermediate runs at other X-levels had been made, and provided drifts in the response level had not occurred, genuine repeat runs would be obtained. With this in mind, repeat runs that show remarkable agreement which is contrary to expectation should always be regarded cautiously and subjected to additional investigation.

Calculation of Pure Error and Lack of Fit Mean Squares

When there are repeat runs in the data, we need additional notation to take care of the multiple observations on Y at the same value of X. Suppose we have m different values of X and, at the jth of these m particular values, X_j, where $j = 1, 2, \ldots, m$, there are n_j observations; we say that:

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are n_1 repeat observations at X_1;
$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are n_2 repeat observations at X_2;
. . .
Y_{ju} is the uth observation ($u = 1, 2, \ldots, n_j$) at X_j;
. . .
$Y_{m1}, Y_{m2}, \ldots, Y_{mn_m}$, are n_m repeat observations at X_m.

Altogether, there are

$$n = \sum_{j=1}^{m} \sum_{u=1}^{n_j} 1 = \sum_{j=1}^{m} n_j$$

observations. The contribution to the pure error sum of squares from the n_1 observations at X_1 is the internal sum of squares of the Y_{1u} about their average $\overline{Y}_1$; that is,

$$\begin{aligned} \sum_{u=1}^{n_1} (Y_{1u} - \overline{Y}_1)^2 &= \sum_{u=1}^{n_1} Y_{1u}^2 - n_1 \overline{Y}_1^2 \\ &= \sum_{u=1}^{n_1} Y_{1u}^2 - \frac{1}{n_1} \left(\sum_{u=1}^{n_1} Y_{1u} \right)^2. \end{aligned} \tag{2.1.2}$$

Provided we are sure that the pure error variation is of the same order of magnitude throughout the data (see Sections 2.2 and 2.3) we next pool the internal sums of squares from all the sites with repeat runs to obtain the overall pure error SS as

$$\sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2 \tag{2.1.3}$$

with degrees of freedom

$$n_e = \sum_{j=1}^{m} (n_j - 1) = \sum_{j=1}^{m} n_j - m. \tag{2.1.4}$$

Thus the pure error mean square is

$$s_e^2 = \frac{\sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2}{\sum_{j=1}^{m} n_j - m} \tag{2.1.5}$$

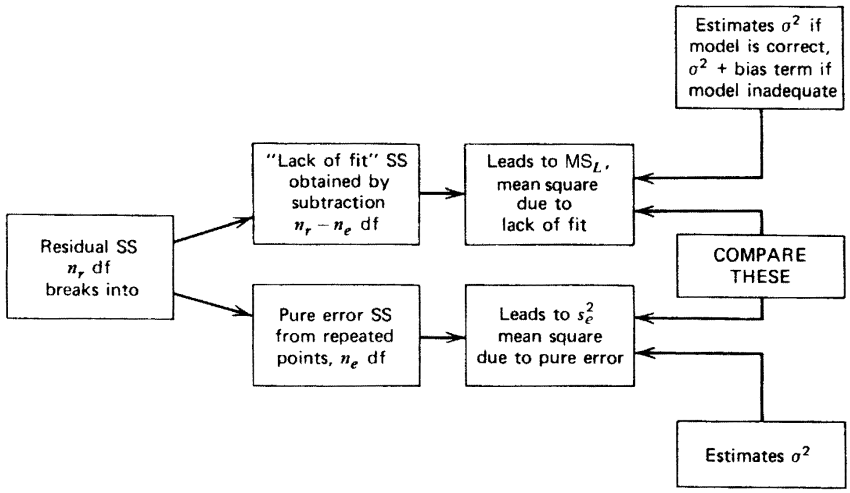


Figure 2.1. Breakup of residual sum of squares into lack of fit and pure error sums of squares.

and is an estimate of σ^2 irrespective of whether the model being fitted is correct or not. In words this quantity is the total of the "within repeats" sums of squares divided by the total of the corresponding degrees of freedom.

Special Formula when $n_j = 2$

If there are only two observations Y_{j1}, Y_{j2} at the point X_j, then

$$\sum_{u=1}^{2} (Y_{ju} - \overline{Y}_j)^2 = \tfrac{1}{2}(Y_{j1} - Y_{j2})^2 \tag{2.1.6}$$

and this is an easier form to compute. This SS has one degree of freedom.

Split of the Residual SS

Now the pure error sum of squares is actually part of the residual sum of squares as we now show. We can write the residual for the uth observation at X_j as

$$Y_{ju} - \hat{Y}_j = (Y_{ju} - \overline{Y}_j) - (\hat{Y}_j - \overline{Y}_j), \tag{2.1.7}$$

using the fact that all the repeat points at any X_j will have the *same* predicted value $\hat{Y}_j$. If we square both sides and sum over both u and j, we obtain

$$\sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \hat{Y}_j)^2 = \sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2 + \sum_{j=1}^{m} n_j(\hat{Y}_j - \overline{Y}_j)^2, \tag{2.1.8}$$

the cross-product vanishing in the summation over u for each j. The left-hand side of Eq. (2.1.8) is the residual sum of squares; the first term on the right-hand side is the pure error sum of squares. The remainder we call the lack of fit sum of squares. It follows that the pure error sum of squares can be introduced into the analysis of variance table as shown in Figure 2.1. The usual procedure is then to compare the ratio $F = MS_L/s_e^2$ with the $100(1 - \alpha)\%$ point of an F-distribution with $(n_r - n_e)$ and n_e degrees of freedom. If the ratio is:

TABLE 2.1. Twenty-three Observations with Same Repeat Runs

Time Order	Y	X	Time Order	Y	X	Time Order	Y	X
12	2.3	1.3	19	1.7	3.7	3	3.5	5.3
23	1.8	1.3	20	2.8	4.0	6	2.8	5.3
7	2.8	2.0	5	2.8	4.0	10	2.1	5.3
8	1.5	2.0	2	2.2	4.0	4	3.4	5.7
17	2.2	2.7	21	3.2	4.7	9	3.2	6.0
22	3.8	3.3	15	1.9	4.7	13	3.0	6.0
1	1.8	3.3	18	1.8	5.0	14	3.0	6.3
11	3.7	3.7				16	5.9	6.7

1. *Significant.* This indicates that the model appears to be inadequate. Attempts would be made to discover where and how the inadequacy occurs. (See comments on the various residuals plots discussed in Sections 2.3–2.6. Note, however, that the plotting of residuals is a standard technique for all regression analyses, not only those in which lack of fit can be demonstrated by this particular test.)

2. *Not Significant.* This indicates that there appears to be no reason *on the basis of this test* to doubt the adequacy of the model and both pure error and lack of fit mean squares can be used as estimates of σ^2. A pooled estimate of σ^2 can be obtained by recombining the pure error and lack of fit sums of squares into the residual sum of squares and dividing by the residual degrees of freedom n_r to give s^2 = (Residual SS)/n_r. (Note that the residuals should *still* be examined because there are other aspects of the residuals to be checked.)

We discussed earlier the fact that repeat runs must be genuine repeats. If they are not genuine repeats, s_e^2 will tend to underestimate σ^2, and the lack of fit F-test will tend to wrongly "detect" nonexistent lack of fit.

Example. Since our previous example, which involved data taken from Appendix 1A, did not contain repeat observations, we shall employ a specially constructed example (see Table 2.1) to illustrate the lack of fit and pure error calculations. A regression line $Y = 1.426 + 0.316X$ was estimated from the data in Table 2.1. The analysis of variance table is shown in Table 2.2. Note that the F-value for regression is not checked *at this stage* because we do not yet know if the model suffers from lack of fit or not.

We now find the pure error, and hence the lack of fit.

1. Pure error SS from repeats at $X = 1.3$ is $\frac{1}{2}(2.3 - 1.8)^2 = 0.125$, with 1 degree of freedom.

2. Pure error SS from repeats at $X = 4.0$ is

TABLE 2.2. ANOVA Table for the Data of Table 2.1

Source	df	SS	MS	F-Ratio
Regression	1	5.499	5.499	7.56 significant at $\alpha = 0.05$ level if no lack of fit
Residual	21	15.278	$0.728 = s^2$	
Total, corrected	22	20.777		

$$(2.8)^2 + (2.8)^2 + (2.2)^2 - 3\{(2.8 + 2.8 + 2.2)/3\}^2$$
$$= 20.52 - (7.8)^2/3$$
$$= 20.52 - 20.28$$
$$= 0.24, \quad \text{with 2 degrees of freedom.}$$

Similar calculations provide the following quantities;

Level of X	$\Sigma_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2$	df
1.3	0.125	1
2.0	0.845	1
3.3	2.000	1
3.7	2.000	1
4.0	0.240	2
4.7	0.845	1
5.3	0.980	2
6.0	0.020	1
Totals	7.055	10

We can thus rewrite the analysis of variance as shown in Table 2.3. The F-ratio $MS_L/s_e^2 = 1.061$ is not significant. Thus, on the basis of this test at least, we have no reason to doubt the adequacy of our model and can use $s^2 = 0.728$ as an estimate of σ^2, in order to carry out an F-test for significance of the overall regression. This latter F-test is valid only if no lack of fit is exhibited by the model and if no other violation of the regression assumptions is apparent. To emphasize this point we summarize the steps to be taken when our data contain repeat observations:

1. Fit the model, write down the usual analysis of variance table with regression and residual entries. Do not perform an F-test for overall regression yet.

2. Work out the pure error sum of squares and break up the residual as in Figure 2.1. (If there is no pure error, lack of fit has to be checked via residuals plots instead; see Sections 2.3–2.6 and Chapter 8.)

3. Perform the F-test for lack of fit. If significant lack of fit is exhibited, go to step 4a. If the lack of fit test is not significant, so that there is no reason to doubt the adequacy of the model, go to step 4b.

4a. Significant lack of fit. Stop the analysis of the model fitted and seek ways to improve the model by examining residuals. Do *not* carry out the F-test for overall regression, and do not attempt to obtain confidence intervals. The assumptions on which these calculations are based are not true if there is lack of fit in the model fitted. (See Section 10.2.)

TABLE 2.3. ANOVA (Showing Lack of Fit Calculation)

Source	df	SS	MS	F-Ratio
Regression	1	5.499	5.499	7.56 significant at $\alpha = 0.05$
Residual	21	15.278	$0.728 = s^2$	
Lack of fit	11	8.233	$0.748 = MS_L$	1.061 (not significant)
Pure error	10	7.055	$0.706 = s_e^2$	
Total, corrected	22	20.777		

4b. Lack of fit test *not* significant. Examine the residuals to see if any other violations of assumptions come to light. If not, recombine the pure error and lack of fit sums of squares into the residual sum of squares, use the residual mean square s^2 as an estimate of $V(Y) = \sigma^2$, carry out an F-test for overall regression, obtain confidence bands for the true mean value of Y, evaluate R^2, and so on.

Note that the fact that the model passes the lack of fit test does not mean that it is *the* correct model—merely that it is a plausible one that has not been found inadequate by the data so far. If lack of fit had been found, a different model would have been necessary—perhaps (here) the quadratic one $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$. Even though the model in our example does not exhibit lack of fit, and has a statistically significant F for overall regression, it is nevertheless not very useful. The R^2 value is only $R^2 = 5.4992/20.7774 = 0.2647$, so that only a small proportion of the variation around $\overline{Y}$ is explained. However, even this pessimistic view of R^2 has to be modified slightly as we now describe.

Effect of Repeat Runs on R^2

As we have already remarked in Section 1.3, it is impossible for R^2 to attain 1 when repeat runs exist, no matter how many terms are used in the model. (A trivial exception is when $s_e^2 = 0$, which rarely happens in practice when there are repeat runs.) No model can pick up the pure error variation (see the solution to Exercise M in "Exercises for Chapters 1–3.")

To illustrate this in our most recent example, we note that the pure error sum of squares is 7.055 with 10 degrees of freedom. No matter what model is fitted to these data, this 7.055 will remain unchanged and unexplained. Thus the maximum R^2 attainable with these data is

$$\text{Max } R^2 = \frac{\text{Total SS, corrected} - \text{Pure error SS}}{\text{Total SS, corrected}}$$

$$= \frac{20.777 - 7.055}{20.777}$$

$$= 0.6604.$$

The value of R^2 actually attained by the fitted model, however, is 0.2674. In other words, we have explained $0.2674/0.6604 = 0.4049$, or about 40% of the amount that can be explained. This figure, while still not too encouraging, looks slightly better. Such a calculation often gives a better sense of what the model actually is achieving in terms of what can be achieved.

Looking at the Data and Fitted Model

The data and the fitted model $\hat{Y} = 1.426 + 0.316X$ are shown in Figure 2.2. We see clearly that, overall, the variation of the points off the line is comparable to the variation within sets of repeats, as already shown by our test for lack of fit with F-ratio slightly over 1. We notice, however, a possible defect not picked up by the lack of fit test. The last observation $(X, Y) = (6.7, 5.9)$ looks a bit remote both from the data and from the line. Clearly other checks are needed to be able to detect this, particularly in larger regressions with several predictors (X's), where a simple plot is not feasible. We shall get to this, and other possible defects, in Chapters 7 and 8. We first finish off our discussion of pure error.

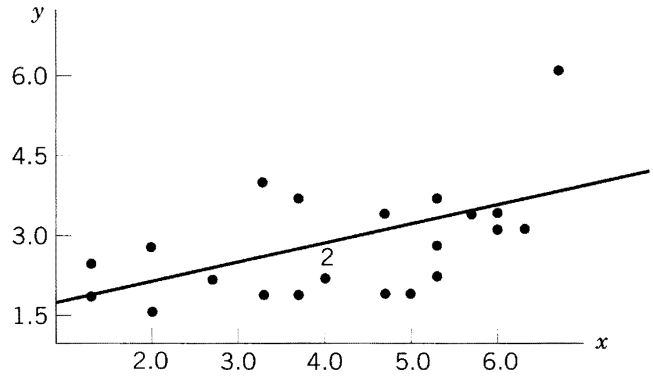


Figure 2.2. Plot of Table 2.1 data and fitted line.

Pure Error in the Many Predictors Case

The formulas given above in the single predictor context apply generally no matter how many predictor variables, $X_1, X_2, \ldots$, are in the data. The only point to watch is that a set of repeat runs must all have the same X_1 value, the same X_2 value, and so on. For example, the four responses at the four points

$$(X_1, X_2, X_3, X_4) = (4, 2, 17, 1), (4, 2, 17, 1), (4, 2, 17, 1), (4, 2, 17, 1)$$

provide repeat runs; however, the four responses at the four points

$$(X_1, X_2, X_3, X_4) = (4, 2, 7, 1), (4, 2, 16, 1), (4, 2, 18, 1), (4, 2, 29, 1)$$

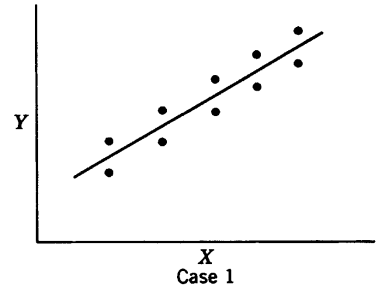
do not, for example, because their X_3 coordinates are all different.

Adding (or Dropping) *X*'s Can Affect Maximum R^2

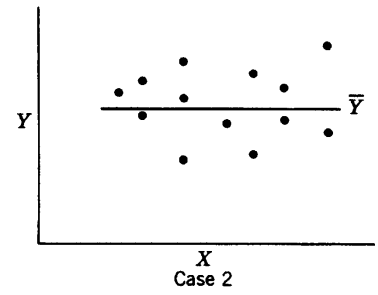
Note that, if additional predictor variables are added to the model, the maximum R^2 value may increase. This is because observations that were repeats before may not be repeats when the additional predictor(s) are introduced. For comments on the treatment of pure error when predictors are *dropped* from the model, see Section 12.2. Dropping predictors can create (pseudo) pure error. An eye needs to be kept on the effects, on the pure error calculation, of all changes of these types.

Approximate Repeats

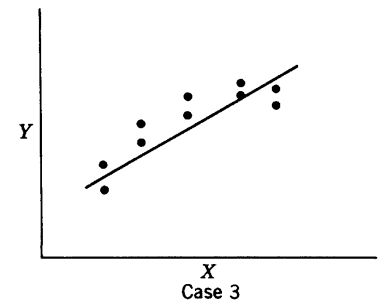
Some sets of data have no, or very few, repeat runs but do have *approximate repeats*, that is, sets of runs that are close together in the X-space compared with the general spread of the points in the X-space. In such cases, we can often use these pseudo-repeats as though they were repeat runs and evaluate an approximate pure error sum of squares from them. This is then incorporated in the analysis in the usual way. For an example of such a use, see Exercise L, in "Exercises for Chapters 1–3." The major problem here is in deciding what the words "close together" mean.



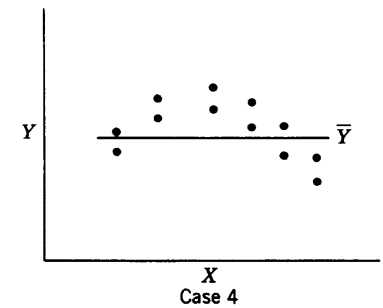
Case 1:
(1) Try $Y = \beta_0 + \beta_1 X + \varepsilon$.
(2) No lack of fit.
(3) Significant linear regression.
(4) Use model $\hat{Y} = b_0 + b_1 X$.



Case 2:
(1) Try $Y = \beta_0 + \beta_1 X + \varepsilon$.
(2) No lack of fit.
(3) Linear regression not significant.
(4) Use model $\hat{Y} = \bar{Y}$.



Case 3:
(1) Try $Y = \beta_0 + \beta_1 X + \varepsilon$.
(2) Significant lack of fit.
(3) Try model $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$.



Case 4:
(1) Try $Y = \beta_0 + \beta_1 X + \varepsilon$.
(2) Significant lack of fit.
(3) Try model $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$.
(*Note:* β_1 may be significantly different from zero when residual error term is reduced by taking out $\beta_{11} X^2$. See Chapter 6.)

Figure 2.3. Typical straight line regression situations.

Generic Pure Error Situations Illustrated Via Straight Line Fits

The diagrams in Figure 2.3 illustrate some situations that may arise when a straight line is fitted to data and the consequent action to be taken. All of these situations are obvious in the context of a straight line fit, but they illustrate situations that occur in more general regressions and our comments should be viewed in that light.

Case 1. The model we try shows no lack of fit and we need to use all of the model.

Case 2. The model we try shows no lack of fit but we do not need all of it.

Case 3. The model we try shows lack of fit and a higher-order (or a different) model must be formulated.

Case 4. The model we try shows lack of fit and, moreover, some of the terms in it seem to be too small to be useful. (A test is not valid here because of the lack of fit.) We must formulate a higher order (or a different) model and must not jump to premature conclusions about terms currently in the model. (For more on such difficulties see Chapter 12.)

(The words "a different model" also include the possibility of transforming Y or X, for example, by using ln Y as a response.)

2.2. TESTING HOMOGENEITY OF PURE ERROR

In practice, we most often look at a plot of the spreads of the repeat runs and decide by eye whether or not they look a lot different from one another. Formal tests exist, if really needed, but all have drawbacks. Cochran's test and Hartley's test require the same number of replicates at each site, plus special tables, so we do not discuss these. Bartlett's test is commonly used but is sensitive to non-normality; that is, if the data are non-normal, the validity of the test is greatly affected. Nevertheless, we describe it and a modified version below. We also describe and recommend Levene's test using group medians rather than means, if such a test is desired. (This essentially converts a test of variances into a test of means, which is relatively unaffected by non-normality. The price paid for this is lowered testing power.)

Bartlett's Test

Let $s_1^2, s_2^2, \ldots, s_m^2$ be the estimates of σ^2 from the m groups of repeats with $\nu_1, \nu_2, \ldots, \nu_m$ degrees of freedom, respectively. In terms of previous notation, $\nu_j = n_j - 1$ and

$$s_j^2 = \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2/(n_j - 1). \tag{2.2.1}$$

As before,

$$s_e^2 = (\nu_1 s_1^2 + \nu_2 s_2^2 + \cdots + \nu_m s_m^2)/(\nu_1 + \nu_2 + \cdots + \nu_m), \tag{2.2.2}$$

and we write $\nu = \nu_1 + \nu_2 + \cdots + \nu_m$. The constant C is defined as

$$C = 1 + \{\nu_1^{-1} + \nu_2^{-1} + \cdots + \upsilon_m^{-1} - \nu^{-1}\}/\{3(m-1)\}, \tag{2.2.3}$$

where m is the number of groups with repeat runs. The test statistic is then

$$B = \left\{\nu \ln s_e^2 - \sum_{j=1}^{m} \nu_j \ln s_j^2\right\}/C. \tag{2.2.4}$$

When the variances of the groups are all the same, B is distributed as χ_{m-1}^2 approximately. A significant B value could indicate inhomogeneous variances. It could also indicate non-normality, so it makes sense to actually look at the shapes of plots of the m samples, too.

Example. Consider the data used to illustrate lack of fit in Section 2.1. We have

$$s_e^2 = 7.055/10 = 0.7055,$$

$$C = 1 + \{6(1/1) + 2(1/2) - 1/10\}/\{3(8-1)\} = 1.328571,$$

$$B = \{10 \ln (0.7055) - \ln (0.125) - \ln (0.845) - \cdots$$

$$-2 \ln (0.980) - \ln (0.020)\}/1.328571$$

$$= \{-3.488485 + 7.836646\}/1.328571 = 3.273(7 \text{ df}).$$

The value of the statistic is very small, indicating no reason to doubt homogeneity of variances. (For example, the 0.95 percentage point of χ_7^2 is 14.1.)

Bartlett's Test Modified for Kurtosis

In this variation, the statistic B of Eq. (2.2.4) is multiplied by $d = 2/(\hat{\beta} - 1)$, where

$$\hat{\beta} = N \sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^4 / \left\{ \sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \overline{Y}_j)^2 \right\}^2 \tag{2.2.5}$$

estimates the (assumed common) kurtosis of the sets of repeats. For normally distributed data the true β would be 3 and d would typically be close to 1. The same χ^2-test as before is used for this statistic; here N is the total number of observations in the (usually reduced) data set used for the test, that is, the total number of observations in all the sets of repeats, ignoring all the single observations in the data.

Example *(Continued)*

$$\hat{\beta} = 18\{[0.25^4 + (-0.25)^4] + \cdots + [0.1^4 + (-0.1)^4]\}/(7.055)^2$$

$$= 18\{5.231\}/(7.055)^2 = 1.891761$$

$$d = 2/(0.891761) = 2.242753$$

$$Bd = 3.273 \times 2.242753 = 7.3405 \quad (7 \text{ df}).$$

The modified test statistic remains nonsignificant compared with $\chi_7^2(0.95) = 14.1$.

Levene's Test Using Means

Consider, in the jth group of repeats, the absolute deviations

$$z_{ju} = |Y_{ju} - \overline{Y}_j|, \qquad u = 1, 2, \ldots, n_j, \tag{2.2.6}$$

of the Y's from the means of their repeats group. Consider this as a one-way classification and compare the "between groups" mean square with the "within groups" mean square via an F-test. The appropriate F-statistic is then

$$\frac{\sum_{j=1}^{m} n_j (\overline{z}_j - \overline{z})^2/(m-1)}{\sum_{j=1}^{m} \sum_{u=1}^{n_j} (z_{ju} - \overline{z}_j)^2 / \sum_{j=1}^{m} (n_j - 1)}, \tag{2.2.7}$$

TABLE 2.4. Details for Levene's Test Using Means for the Data of Table 2.1

X-Level, X_j	z_{ju}	n_j	$\bar{z}_j$
1.3	0.25, 0.25	2	0.25
2.0	0.65, 0.65	2	0.65
3.3	1.00, 1.00	2	1.00
3.7	1.00, 1.00	2	1.00
4.0	0.20, 0.20, 0.40	3	$0.2\dot{6}$
4.7	0.65, 0.65	2	0.65
5.3	0.00, 0.70, 0.70	3	$0.4\dot{6}$
6.0	0.10, 0.10	2	0.10

where

$$\bar{z}_j = \sum_{u=1}^{n_j} z_{ju}/n_j, \qquad \bar{z} = \sum_{j=1}^{m}\sum_{u=1}^{n_j} z_{ju}/\sum_{j=1}^{m} n_j. \tag{2.2.8}$$

The F-value is referred to $F\{m - 1, \Sigma_{j=1}^{m}(n_j - 1)\}$, using only the upper tail.

Example *(Continued)*. We have $m = 8$, $\Sigma_{j=1}^{m} n_j = 18$, $\Sigma_{j=1}^{m}(n_j - 1) = 10$, $\bar{z} = 9.5/18 = 0.52\dot{7}$. The z_{ju} and the row means $\bar{z}_j$ are shown in Table 2.4, where we use only the repeat runs, ignoring the singles, which do not contribute here.

The numerator is then $1.687783/(8 - 1) = 0.24112$ and the denominator is $0.353333/10 = 0.035333$, whereupon $F = 6.824$, which we can compare to $F(7, 10, 0.95) = 3.14$. This would indicate that there *are* differences between the variances of the various groups. We comment on this further below.

Levene's Test Using Medians

Consider, in the jth group of repeats, the absolute deviations

$$z_{ju} = |Y_{ju} - \tilde{Y}_j|, \qquad u = 1, 2, \ldots, n_j,$$

of the Y's from the *medians* $\tilde{Y}_j$ of their repeats group. Consider these in a one-way classification and compare the "between groups" mean square with the "within groups" mean square via an F-test. The appropriate F-statistic is again (2.2.7), and it is tested in the same way as before. See Carroll and Schneider (1985).

Example *(Continued)*. Note that when only two observations are in a group, the mean and median are identical. So only groups with three or more observations can give z_{ju} and $\bar{z}_j$ values different from those in Table 2.4. For $X = 4.0$, the median is 2.8; the mean was 2.6. For $X = 5.3$ the median is 2.8, identical to the mean, as it happens. So the F-statistic changes only slightly *for this example* through the changed calculation for $X = 4.0$; the new z_{ju} values there are 0, 0, and 0.6 with $\bar{z}_j = 0.2$ (replacing 0.2, 0.2, and 0.4 with mean $0.2\dot{6}$). Now $\bar{z} = 9.3/18 = 0.51\dot{6}$. This gives $F = \{1.803333/(8 - 1)\}/\{0.566667/10\} = 4.546$, smaller than in the foregoing test, but still greater than $F(7, 10, 0.95) = 3.14$. So again differences are declared between the variances of the groups.

Some Cautionary Remarks

Our numerical example is (on the one hand) not a particularly good one to illustrate the Levene tests because the denominator of the test statistic is estimated by only

two sets of three z_{ju} values; the pairs do not contribute to the within sum of squares. On the other hand, it does alert us to such possible problems! It is also worrying that, although the pairs do not contribute to the within groups numerical value, they are granted a degree of freedom! Alan Miller has suggested a sensible possible adjustment, reducing these df to zero, but this does not seem to solve the problem either. Simulations performed by T.-S. Lim and W.-Y. Loh, some of which are mentioned in Lim and Loh (1996) and some of which were performed privately as a favor to the authors of this book, seem to indicate that the best test is Levene's test using medians. (Our example would indicate that the data should not contain too many *pairs* of repeats, however.) At the same time, it makes practical sense to plot the Y-values and visually to compare the repeat groupings with one another. So that is our somewhat cautious joint recommendation, with the plots always taking preference. (In using these example data again later, we do not make any adjustments for possible unequal variances, since the evidence for this seems weak.)

A Second Example

The groups of data below are values from our Exercise 23D, adapted via $(Y - 1430)/5$.

Group:	1	2	3	4	5
	52	24	39	59	20
	30	3	4	24	23
	63	43	16	0	27
	51	23	—	3	18

The Bartlett test value is, from (2.2.4), $B = 6.83$. Adjusting via (2.2.5) leads to $Bd = 6.41$. Both values are less than $\chi^2_{4,0.95} = 9.49$. The F-statistics from Levene's tests are 1.56 (using means) and 1.16 (using medians). Both are less than $F_{4,14,0.95} = 3.11$. So in this example we have consistent conclusions *not* rejecting homogeneity.

2.3. EXAMINING RESIDUALS: THE BASIC PLOTS

As we have already mentioned, the residuals $e_i = Y_i - \hat{Y}_i$ contain within them information on why the model might not fit the data. So it is well worthwhile to check the behavior of the residuals and allow them to tell us of any peculiarities of the regression fit that might have occurred.

The study of residuals is not new, as the following quotation makes clear.

> Almost all the greatest discoveries in astronomy have resulted from the consideration of what we have elsewhere termed RESIDUAL PHENOMENA, of a quantitative or numerical kind, that is to say, of such portions of the numerical or quantitative results of observation as remain outstanding and unaccounted for after subducting and allowing for all that would result from the strict application of known principles. (Sir John F. W. Herschel, Bart. K. H., 1849, p. 548)

An enormous amount has been written about the study of residuals. There are, in fact, several excellent books (see Section 2.8). In this section we discuss only the basic plots that allow the most useful checks. These are the checks that should be done on a routine basis for every regression. More sophisticated methods are discussed in later chapters for those wishing to explore further.

The work of this section is useful and valid not only for linear regression models

but also for nonlinear regression models and analysis of variance models. In fact, this section applies to *any* situation where a model is fitted and measures of unexplained variation (in the form of a set of residuals) are available for examination. Thus, like the pure error calculations in Section 2.1, it is *not* restricted only to the straight line regression case, even though we find it convenient to talk about it here.

How Should the Residuals Behave?

The residuals are defined as the n differences $e_i = Y_i - \hat{Y}_i$, $i = 1, 2, \ldots, n$, where Y_i is an observation and $\hat{Y}_i$ is the corresponding fitted value obtained by use of the fitted regression equation.

Note: Usually, the residuals would be evaluated to the same number of significant figures as appeared in the original response observations. Sometimes one additional significant figure is used, but to go beyond this is generally a waste of effort. Computer printouts typically contain more figures than necessary, of course, but these would be cut back if the data were transcribed for reporting purposes.

We can see from their definition that the residuals e_i are the differences between what is actually observed, and what is predicted by the regression equation—that is, the amount that the regression equation has not been able to *explain.* Thus we can think of the e_i as the *observed errors if the model is correct.* (There are, however, restrictions on the e_i induced by the normal equations.) Now in performing the regression analysis we have made certain assumptions about the errors; the usual assumptions are that the errors are independent, have zero mean, have a constant variance σ^2, and follow a normal distribution. The last assumption is required for making F-tests. Thus if our fitted model is correct, the residuals should exhibit tendencies that tend to confirm the assumptions we have made or, at least, should not exhibit a denial of the assumptions. This latter idea is the one that should be kept in mind when examining the residuals. We should ask: "Do the residuals make it appear that our assumptions are wrong?" After we have examined the residuals we shall be able to conclude either that (1) the assumptions appear to be violated (in a way that can be specified) or (2) the assumptions do not appear to be violated. Note that (2) does not mean that we are concluding that the assumptions are correct; it means merely that on the basis of the data we have seen, we have no reason to say that they are incorrect. The same spirit occurs in making tests of hypotheses when we either *reject* or *do not reject* (rather than *accept*). We now give ways of examining the residuals in order to check the model. These ways are all graphical, are easy to do, and are usually very revealing when the assumptions are violated. The principal ways of plotting the residuals e_i are:

1. To check for non-normality.
2. To check for time effects if the time order of the data is known.
3. To check for nonconstant variance and the possible need for a transformation on Y.
4. To check for curvature of higher order than fitted in the X's.

In addition to these basic plots, the residuals should also be plotted:

5. In any way that is sensible for the particular problem under consideration.

(We remark before proceeding that the basic plots should always be done and will often pick up any deficiencies present in many sets of residuals. It is also possible for these simple plots to be "fooled" or "defeated," however, if a sophisticated defect, or a combination of defects, occurs. That is why more complicated methods of analyzing

residuals have been developed. The methods of this chapter are, however, the first line of defense for detection of an unsuitable model.)

We talk here, and continue to do so in this chapter, of looking at the ordinary residuals, defined as $Y_i - \hat{Y}_i$. Actually there are several types of residuals, any or all of which could be obtained and could be plotted as well, or instead. We discuss the various types in Chapter 8. For most regressions, it is not crucial which set of residuals is plotted. Occasionally it makes a great deal of difference which choice of residuals is made; this would show up when the plots of various residual sets are compared.

2.4. NON-NORMALITY CHECKS ON RESIDUALS

We usually assume that $\epsilon_i \sim N(0, \sigma^2)$ and that all errors are independent of one another. Their estimates, the residuals, cannot be independent. The estimation of the parameters (p of them, say; $p = 2$ for the straight line) means that the n residuals carry only $(n - p)$ df. The p normal equations [for $p = 2$, see Eqs. (1.2.8)] are restrictions on the e_i, essentially. Unless p is large compared with n, this typically has little effect on our non-normality checks. We first note that:

For any model with a β_0 (intercept) term in it, the least squares residuals must, in theory, add to zero.

This is seen from the first normal equation obtained by differentiating the error sum of squares with respect to β_0. If the model fitted is $E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, the equation can be written

$$-2\,\Sigma(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki}) = 0,$$

where the summation is taken over $i = 1, 2, \ldots, n$. This reduces to

$$\Sigma(Y_i - \hat{Y}_i) = 0.$$

Thus

$$\Sigma\, e_i = 0.$$

Because the least squares fitting procedure guarantees this, there is no need to check that the mean $\bar{e} = \Sigma e_i/n$ is zero. We have made it so!

Often a simple histogram, or a stem and leaf plot, will be enough. Figure 2.4 shows histograms for the residuals from (*a*) the steam data fit of Chapter 1 and (*b*) the straight line fit for the lack of fit test data in Section 2.1. We conclude that these (somewhat crude) plots look "normal enough," sometimes a difficult judgment, except for the highest observation in Figure 2.4*b*, which looks like an outlier.

Normal Plot of Residuals

An alternative and (we believe) better check is to make a *normal probability plot.* This is not difficult to do "by hand" but is better done in the computer. A full explanation is given in Appendix 2A. Here we merely set out the steps required in the MINITAB system of computing. Find out which column contains the residuals you wish to "normal plot", say, c11. Write

```
nscore c11  c12
plot c12  c11
```

Draw (or imagine) a straight line through the main middle bulk of the plot. Ask:

Midpoint	Count	
−1.6	1	*
−1.2	3	***
−0.8	1	*
−0.4	4	****
0.0	7	*******
0.4	2	**
0.8	3	***
1.2	4	****

(*a*) Histogram of steam residuals $n = 25$

Midpoint	Count	
−1.0	4	****
−0.5	6	******
0.0	6	******
0.5	4	****
1.0	1	*
1.5	1	*
2.0	0	
2.5	1	*

(*b*) Histogram of residuals $n = 23$

Figure 2.4. Histograms of residuals from (*a*) steam data fit and (*b*) lack of fit test data.

"Do all the points lie on such a line, more or less?" If the answer is yes, one would conclude that the residuals do not deny the assumption of "normality of errors" made in performing tests and getting confidence intervals. We see from the plots in Figures 2A.4, 2A.5 and 2A.6 that the steam data residuals look alright, but that the pure error example data show signs of an *outlier*, an observation that falls unusually out of the pattern for a normal sample. For *why* it is an outlier, see Appendix 2A.

(*Note carefully:* If the plot is made "the other way around," that is, as

```
nscore c11  c12
plot c11  c12
```

the criteria for an outlier to exist are different. It follows that every normal plot must be looked at carefully to see which axis is which, before conclusions are reached. This step is *very* important.)

There are other ways of assessing normality. For information on the Shapiro and Wilk test, a useful starter reference is Royston (1995).

2.5. CHECKS FOR TIME EFFECTS, NONCONSTANT VARIANCE, NEED FOR TRANSFORMATION, AND CURVATURE

We plot the residuals e_i vertically against, in turn:

1. The time order of the data, if known.
2. The corresponding fitted values $\hat{Y}_i$, using the fitted model.
3. The corresponding X_i values if there is only one predictor variable; or, in general, each set of X_{ji}, where $j = 1, 2, \ldots, k$ represent the X's in the regression.

Figure 2.5. A satisfactory residuals plot should give this overall impression.

In all of these cases, a satisfactory plot is one that shows a (more or less) horizontal band of points giving the impression of Figure 2.5. There are many possible unsatisfactory plots. Three typical ones appear in Figure 2.6. The first of these three (the funnel) displays the band of residuals widening to the right showing nonconstant variance. The second is an upward trend and the third is curvature. (All of these defective plots can appear in other directions, of course, for example, a reversed funnel or a downward curve.) It is difficult to be absolutely specific about what to do if these defects are found but Table 2.5 gives some general indications.

Plots for the steam data and the lack of fit example data appear in Figures 2.7 and 2.8. We see that, for the steam data, none of the three plots seems to show any worrying anomaly that would indicate the regression fit is defective. For the lack of fit example data, we must remember that the observation with the largest residual (value 2.36) is a likely outlier. If we ignore this residual in the residuals versus time plot, there is still a hint of a funnel shape, but perhaps too little to act on—opinions would differ on this. The plot of residuals versus $\hat{Y}$ values again shows up the outlier but is otherwise unremarkable. (The apparent slight "downward slope look" caused by ignoring the outlier is essentially "caused by" the presence of the outlier.) The plot of residuals versus X is similar to the foregoing plot, because Y and X rise together. In our two examples, the residuals are equally spaced in time. If they were not, and the correct spacings were known, the residuals would be plotted using those spacings, of course.

Three Questions and Answers

Query 1. Why do we plot the residuals $e_i = Y_i - \hat{Y}_i$ against the $\hat{Y}_i$ and not against the Y_i, for the usual linear model?

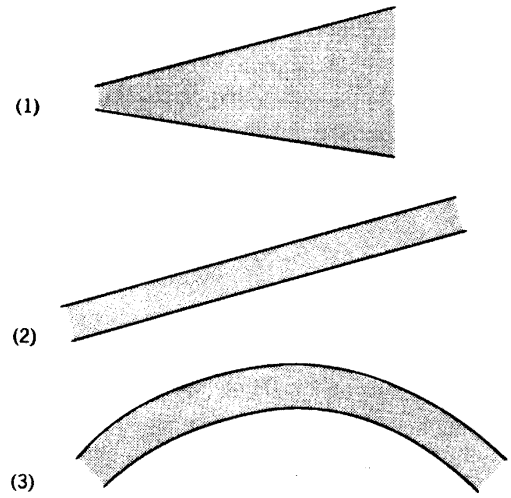


Figure 2.6. Examples of characteristics shown by unsatisfactory residuals behavior.

TABLE 2.5. Possible Remedies for Unsatisfactory Residuals Plots

Unsatisfactory Plot: See Figure 2.6	Plot of e_i Versus		
	Time Order	Fitted $\hat{Y}_i$	X_{ji} Values
Funnel indicating nonconstant variance	Use weighted[a] least squares	Use weighted[a] least squares or transform[b] the Y_i	Use weighted[a] least squares or transform[b] the Y_i
Ascending or descending band	Consider adding first-order term in time	Error in analysis or wrongful omission of β_0	Error in the calculations; first-order effect of X_j not removed
Curved band	Consider adding first- and second-order terms in time	Consider adding extra terms to the model or transform[b] the Y_i	Consider adding extra terms to the model or transform[b] the Y_i

[a]See Section 9.2.
[b]See Chapter 13.

Answer. Because the e's and the Y's are usually correlated but the e's and the $\hat{Y}$'s are not. One way to see this is to think of plots of the e_i as ordinate against (i) the Y_i and (ii) the $\hat{Y}_i$, and find the slope of a least squares lines through the points. For (i) it will be $1 - R^2$; for (ii) 0. This means that, unless $R^2 = 1$, there will always be a slope of $1 - R^2$ in the e_i versus Y_i plot, even if there is nothing wrong. However, a slope in the e_i versus $\hat{Y}_i$ plot *indicates* that something is wrong. See Exercise X in "Exercises for Chapters 5 and 6."

Query 2. Why does the plot of residuals $e_i = Y_i - \hat{Y}_i$ versus $\hat{Y}_i$ exhibit a series of straight lines, with slopes of -1?

Answer. This feature is in fact *always* present but is usually not obvious. A line is formed by any set of plotted points with the same Y value. Suppose, for example, we have m points with the same value of Y, $Y = a$, say. Then we plot this subset of residuals

$$a - \hat{Y}_1, \quad a - \hat{Y}_2, \ldots, \quad a - \hat{Y}_m,$$

$$\text{versus} \qquad \hat{Y}_1, \qquad \hat{Y}_2, \ldots, \qquad \hat{Y}_m, \qquad \text{respectively.}$$

These m points are all on a line with slope -1 through the points $(\hat{Y}, e) = (a, 0)$ and $(0, a)$. For, if $\tilde{Y}$ is the average $\tilde{Y} = (\hat{Y}_1 + \hat{Y}_2 + \cdots + \hat{Y}_m)/m$, the slope of the line, via least squares, is

$$\frac{S_{e\hat{Y}}}{S_{\hat{Y}\hat{Y}}} = \frac{\Sigma(a - \hat{Y}_i - (a - \tilde{Y}))(\hat{Y}_i - \tilde{Y})}{\Sigma(\hat{Y}_i - \tilde{Y})^2} = -1,$$

after cancellation of the a's. The intercept is $(a - \tilde{Y}) - (-1)\tilde{Y} = a$.

In data sets where only a limited number of Y's are recorded (e.g., color levels of a dyestuffs product, percentages of pests present on a plant leaf) this feature may become very obvious. Searle (1988, p. 211) who drew attention to this feature also points out that:

1. The lines always exist. When no Y's are repeated, there is only one point on each line.

2. The lines occur no matter what model is fitted, and whether linear, nonlinear, or generalized linear model estimation techniques have been used.

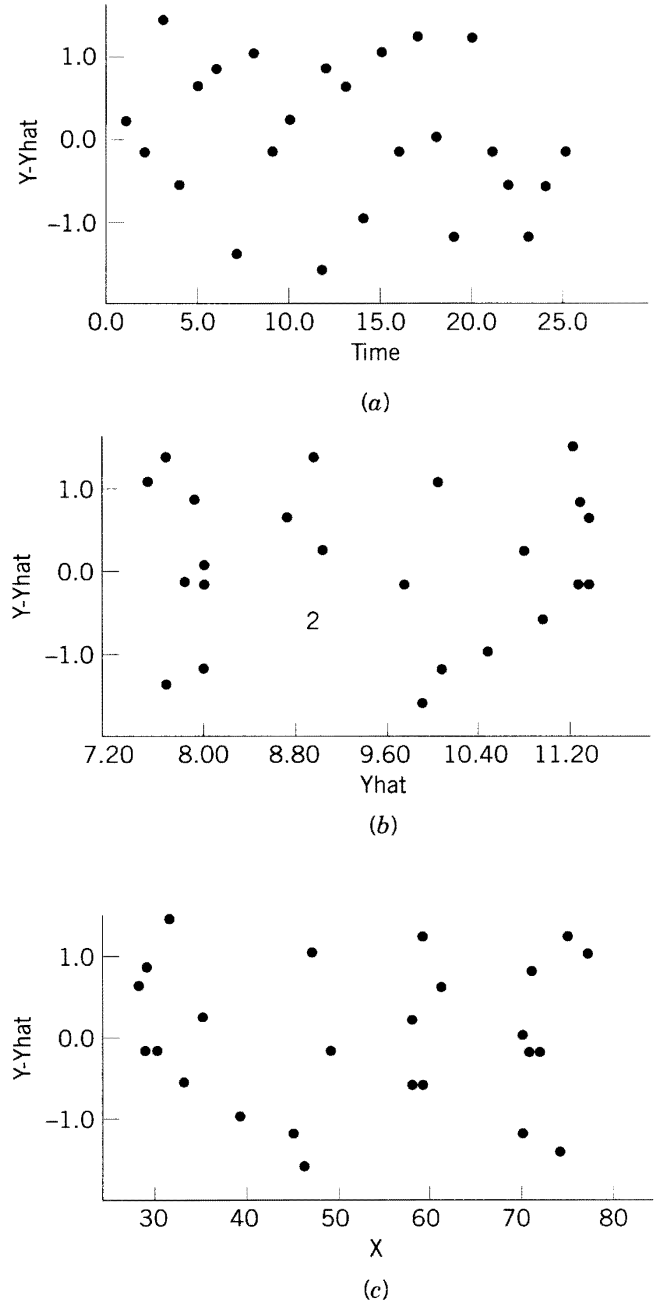


Figure 2.7. Plots of steam data residuals versus (*a*) order, (*b*) $\hat{Y}$, and (*c*) X.

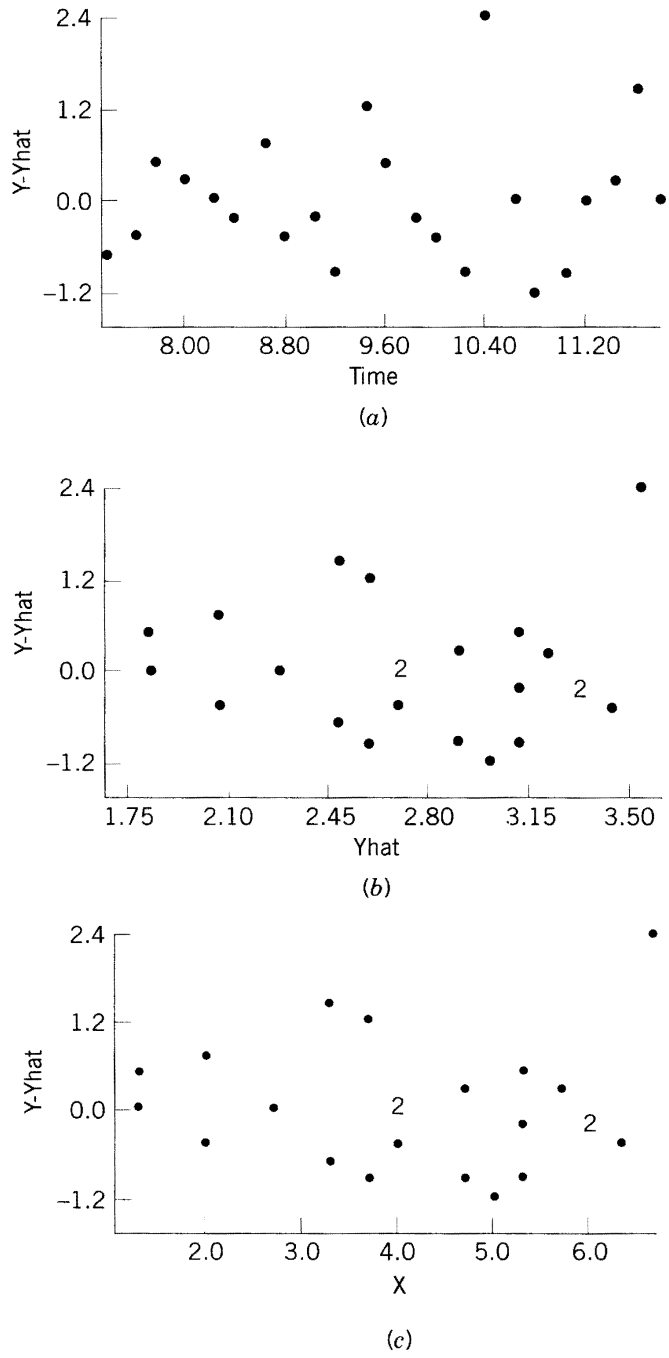


Figure 2.8. Plots of lack of fit data residuals versus (*a*) order, (*b*) $\hat{Y}$, and (*c*) X.

Example. The data in Table 2.1 have four data points with $Y = 2.8$ with residuals from a straight line fit $\hat{Y} = 1.426 + 0.316X$ of:

Number	e	$\hat{Y}$
(7)	$2.8 - 2.05 =$	0.75 plotted against 2.05
(20)	$2.8 - 2.68 =$	0.12 plotted against 2.68
(5)	$2.8 - 2.68 =$	0.12 plotted against 2.68
(6)	$2.8 - 3.10 =$	−0.30 plotted against 3.10

It is easily confirmed that the four points lie on a line with slope -1 and through points (0, 2.8), (2.8, 0), that is, on the line $e + \hat{Y} = 2.8$.

Query 3. Is it possible to work out some test statistics instead of looking at the diagrams?

Answer. It is possible to evaluate test statistics, but it is often difficult to know if they are sufficiently deviant to require action. In practical regression situations, a detailed examination of the corresponding residuals plots is usually far more informative, and the plots will almost certainly reveal any violations of assumptions serious enough to require corrective action.

Consider the plot of e_i against $\hat{Y}_i$ described above. Three particular types of discrepancies were mentioned and related to the diagrams of Figure 2.6. We can measure each of these defects with appropriate statistics as follows. Define

$$T_{pq} = \Sigma_{i=1}^{n} e_i^p \hat{Y}_i^q. \tag{2.5.1}$$

Then:

1. $T_{21} = \Sigma_{i=1}^{n} e_i^2 \hat{Y}_i$ provides a measure for the type of defect shown in Figure 2.6(1).
2. $T_{11} = \Sigma_{i=1}^{n} e_i \hat{Y}_i$. This should always be zero. This provides a measure for the defect shown in Figure 2.6(2). Evaluation of this statistic could be done as a routine check, if desired.
3. $T_{12} = \Sigma_{i=1}^{n} e_i \hat{Y}_i^2$ provides a measure for the type of defect shown in Figure 2.6(3). It is related to Tukey's "one degree of freedom for nonadditivity" statistic. (See also Exercise O in "Exercises for Chapters 5 and 6.")

Other types of statistics are also available. Readers who would like to learn more about them should consult the texts listed in Section 2.8.

Comment

The plots we have discussed are very basic ones and can be criticized in a number of ways because they may not show up defects of specific types. A vast literature has grown up, and many more sophisticated methods have been suggested. We deal with some of these in Chapter 8, and these provide further references to which interested readers can turn after that.

2.6. OTHER RESIDUALS PLOTS

Specialist knowledge of the problem under study often suggests that other types of residuals plots should be examined. For example, suppose it were known that the 23 observations that led to the 23 residuals from the lack of fit test example came

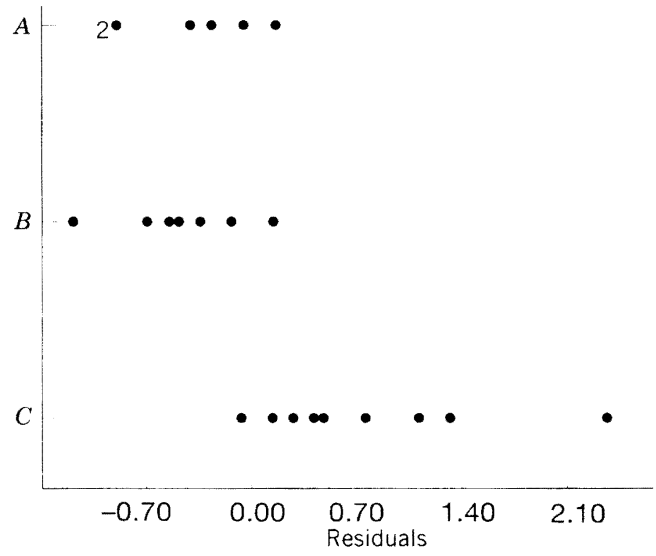


Figure 2.9. Residuals plot indicating block effects not incorporated in the fitted model.

from three separate machines A, B, and C, so that the residuals when grouped by machines were

A: −0.08, −0.89, 0.11, −1.01, −0.30, −1.00, −0.42

B: −0.56, −0.67, 0.11, −0.49, −1.20, −0.12, −0.32

C: 0.46, −0.04, 0.74, 1.33, 1.11, 0.29, 0.40, 0.17, 2.36

(The machine order for the observations in Table 2.1 is thus *CCCBA CBCAA BBCAB CAACB BAC.*)

Figure 2.9 shows a plot against machines. This would suggest that there is a basic difference in level of response Y of machine C compared with A and B. Such a difference could be incorporated into the model by the introduction of a dummy variable; this is discussed in Chapter 14.

Another example of "other residual plots" occurs when a possible new variable comes into consideration. Suppose it is suspected that the ambient temperature is affecting the contents of a large vessel. Although vessel temperature has been recorded at a selected, protected, measuring point, the temperature at the other side of the vessel may possibly be affected by exposure to the outside air. If the ambient temperatures are recorded for the period during which data were collected, the residuals could now be plotted against the temperatures observed, to see if any dependency of response on ambient temperature is revealed. If it is, new terms of appropriate kinds can be added to the model to take account of the dependency.

These are two examples of what "other residuals plots" might be used. In general, residuals should be plotted in *any* reasonable way that occurs to the experimenter or statistician, based on specialist knowledge of the problem under study. The plots already described are, however, the basic ones and should always be performed for a full analysis.

Dependencies Between Residuals

As we have remarked, the residuals, unlike the errors they estimate, are not independent. Does this affect the plots? Yes. Does it invalidate the plots? In most situations,

no. Anscombe and Tukey (1963, p. 144) remark on this point. In discussing the two-way analysis of variance (where there are several constraints on the residuals) they remark that, although correlations and constraints affect distributions of functions of the residuals, the "corresponding effects on the graphical procedures . . . can usually be neglected. This is mainly because of the way in which graphical appearances arise from residuals, though in part because of the absence of precisely defined significance levels. (This is also true for most other balanced designs.)" In a later sentence Anscombe and Tukey state that in a two-way table with four or more rows and four or more columns, "the effect of correlation upon graphic procedures is usually negligible. . . ." It would appear that in general regression situations the effect of correlations between residuals need not be considered when plots are made, except when the ratio $(n - p)/n$—that is, (number of degrees of freedom in residuals)/(number of residuals)—is quite small.

In Chapter 8 we shall see how to evaluate the pairwise correlations between the residuals. If these correlations are relatively small, there is usually little effect on the residuals plots.

2.7. DURBIN–WATSON TEST

We later (Chapter 7) explain the Durbin–Watson test in some detail for multiple predictors. Here we merely sketch its application to the residuals obtained from fitting a straight line $\hat{Y} = b_0 + b_1 X$. It is assumed here that the observations, and so the residuals, have a natural order such as a time order or space order, here indicated by the order $Y_1, Y_2, \ldots, Y_n$. In practice, the given data might have to be recast to obtain the proper ordering. The residuals $e_1, e_2, \ldots, e_n$ are estimates for errors assumed to be independent. If they are not independent, the residuals may reveal it. The Durbin–Watson test checks for a sequential dependence in which each error (and so residual) is correlated with those before and after it in the sequence. The test focuses specifically on the differences between successive residuals in the following way. Consider the Durbin–Watson statistic

$$d = \sum_{u=2}^{n} (e_u - e_{u-1})^2 / \sum_{u=1}^{n} e_u^2. \tag{2.7.1}$$

It can be shown that:

1. $0 \leq d \leq 4$ always.
2. If successive residuals are positively serially correlated, that is, positively correlated in their sequence, d will be near 0.
3. If successive residuals are negatively correlated, d will be near 4, so that $4 - d$ will be near 0.
4. The distribution of d is symmetric about 2.

Because of (4), a $d < 2$ should be used as is; a $d > 2$ should be tested as $4 - d$ and point (3) should be kept in mind. The test is conducted as follows. Compare d (or $4 - d$, whichever is closer to zero) with d_L and d_U in Table 2.6. If $d < d_L$, conclude that positive serial correlation is a possibility; if $d > d_U$, conclude that no serial correlation is indicated. (If $4 - d < d_L$, conclude that negative serial correlation is a possibility; if $4 - d > d_U$, conclude that no serial correlation is indicated.) If the d (or $4 - d$) value lies between d_L and d_U, the test is inconclusive. An indication of

TABLE 2.6. Significance Points of d_L and d_U for a Straight Line Fit

	1%		2.5%		5%	
n^a	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.95	1.23	1.08	1.36
20	0.95	1.15	1.08	1.28	1.20	1.41
25	1.05	1.21	1.18	1.34	1.29	1.45
30	1.13	1.26	1.25	1.38	1.35	1.49
40	1.25	1.34	1.35	1.45	1.44	1.54
50	1.32	1.40	1.42	1.50	1.50	1.59
70	1.43	1.49	1.51	1.57	1.58	1.64
100	1.52	1.56	1.59	1.63	1.65	1.69
150	1.61	1.64	—	—	1.72	1.75
200	1.66	1.68	—	—	1.76	1.78

[a]Interpolate linearly for intermediate n-values.

positive or negative serial correlation would be cause for the model to be reexamined. Weighted, or generalized, least squares (Section 9.2) becomes a possibility.

2.8. REFERENCE BOOKS FOR ANALYSIS OF RESIDUALS

Full reference details of the following are in the bibliography: Atkinson (1985); Barnett and Lewis (1994); Belsley (1991); Belsley, Kuh, and Welsch (1980); Chatterjee and Hadi (1988); Cook and Weisberg (1982); Hawkins (1980); and Rousseeuw and Leroy (1987).

APPENDIX 2A. NORMAL PLOTS

The area under an $N(0, 1)$ distribution from $-\infty$ to some point x is given by

$$y = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right) dt. \tag{2A.1}$$

If we plot $100y$ as ordinate, against x as abscissa, we obtain the "S-shaped" curve, called the cumulative probability curve of the $N(0, 1)$ distribution. Some points on this curve are, for example, $(x, y) = (-1.96, 2.5)$, $(0, 50)$, and $(1.96, 97.5)$, all of which are easily obtained from tables of the cumulative $N(0, 1)$ distribution. (See Figures 2A.1 and 2A.2.)

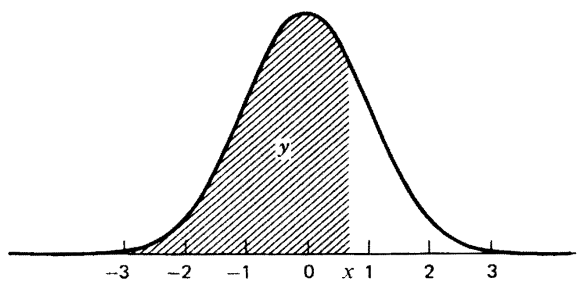


Figure 2A.1. Cumulative area under the normal distribution to point x.

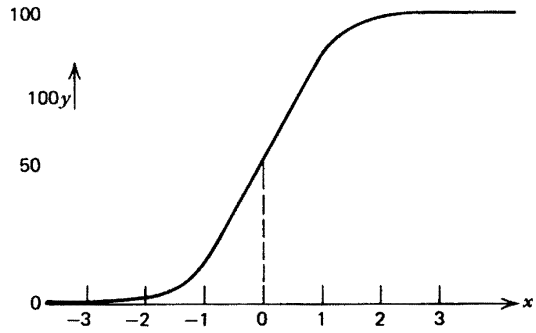


Figure 2A.2. Cumulative normal curve.

Normal probability paper is a specially constructed type of graph paper that is available at most technical bookstores. Although the unnumbered horizontal axis is marked by equal divisions in the usual way, the vertical axis has a special scale. The vertical scale goes from 0.01 to 99.99 but the spacing of the divisions becomes wider as we move up from the 50 point to the 99.99 point and down from the 50 point to the 0.01 point, with symmetry about the horizontal 50 line. The scaling is such that if 100 times the value of y in (2A.1) is plotted against x, the resulting "curve" will be a straight line. Thus the vertical scaling, determined from the inverse functions of Eq. (2A.1), $x = F^{-1}(y)$, "straightens out" the top and bottom of the S-shaped curve in Figure 2A.2. Note that since the points $(-\infty, 0)$ and $(\infty, 100)$ are on the straight line plot, the values 0 and 100 cannot be shown on the scale since the horizontal scale is of limited length and cannot go from $-\infty$ to ∞. A further point on the straight line mentioned above is (1, 84.13). We shall find this point useful in a moment.

If points from the cumulative $N(\mu, \theta^2)$ distribution are plotted on normal probability paper [rather than points from the $N(0, 1)$ distribution], then a straight line will pass through such points as $(x, y) = (\mu - 1.96\theta, 2.5)$, $(\mu, 50)$, $(\mu + \theta, 84.13)$, $(\mu + 1.96\theta, 97.5)$, and so on. This fact is very useful if we have a sample $x_1, x_2, \ldots, x_m$ and wish to decide if it could have come from a normal distribution, and if so, to obtain a quick estimate of the standard deviation θ. First, the sample is arranged in ascending order, due regard being given to sign. Let us assume this has been done already so that $x_1, x_2, \ldots, x_m$ is the correct order. We now plot x_i against the ordinate[1] with value

$$100(i - \tfrac{1}{2})/m. \tag{2A.2}$$

The rationale behind this is that, if we divide the unit area under the normal curve into m equal areas, we might "expect" that one observation lies in each section so marked out. Thus the ith observation in order, x_i, is plotted against the cumulative area to the middle of the ith section, which is $(i - \frac{1}{2})/m$. The factor 100 adapts this to the vertical scale given on the normal probability paper. (See Figure 2A.3.)

If the sample is a normal sample it will be found that a well-fitting straight line can be drawn (by eye) through the bulk of the points plotted, although none of the points may necessarily fall right on the line. We can then use the best-fitting straight line to estimate θ as follows. Find x_{50} and $x_{84.13}$, the values of x for which the line crosses

[1]For possible alternatives to $100(i - \frac{1}{2})/m$, see Barnett (1975); note especially the last paragraph of p. 101 and the first paragraph of p. 104. The BMDP programs use $100(3i - 1)/(3m + 1) = 100(i - \frac{1}{3})(m + \frac{1}{3})$ and also produce a "detrended normal probability plot" from which the slope has been removed. MINITAB uses $100(i - \frac{3}{8})/(m + \frac{1}{4})$ and converts this to a normal score. The differences between these different systems are typically unimportant in practical use.

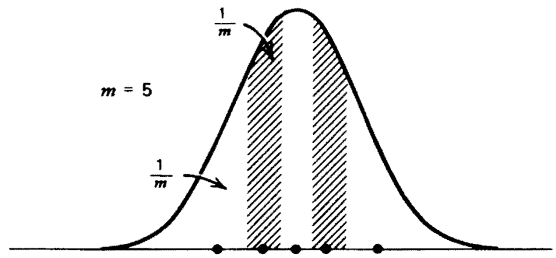


Figure 2A.3. Splitting the area under the normal curve into m equal pieces; we might "expect" one observation in each piece at a location that divides the area of the piece into two equal portions.

horizontal lines drawn at 50 and 84.13 ordinate levels. Then the difference $x_{84.13} - x_{50}$ is an estimate of $[(\mu + \theta) - \mu] = \theta$. (See Figure 2A.4 write-up below.)

An instructive way to gain experience to make decisions on these types of plots is to look up samples of various sizes from a table of random normal deviates and to plot them on normal probability paper. This will give an idea of the variation from linearity that *can* occur and that is *not* abnormal. Plots of this type are given by Daniel and Wood (1980, Appendix 3A).

Normal Scores

Most normal plots are done on the computer, and there the vertical axis is often converted to a *normal score*, that is, the normal deviate value that would correspond to the plotted probability level. (For example, 2.28% would be converted to −2, 2.5% to −1.96, 50% to zero, and 99.865% to 3. See the normal probability table.) In the MINITAB system, this is particularly simple to achieve, If the residuals are in column C6, we write

```
nscore c6  c7
plot c7  c6
```

and the plot is made. Note that the plot instruction must be written in that way to get the diagram to look similar to our discussion and examples below. Use of

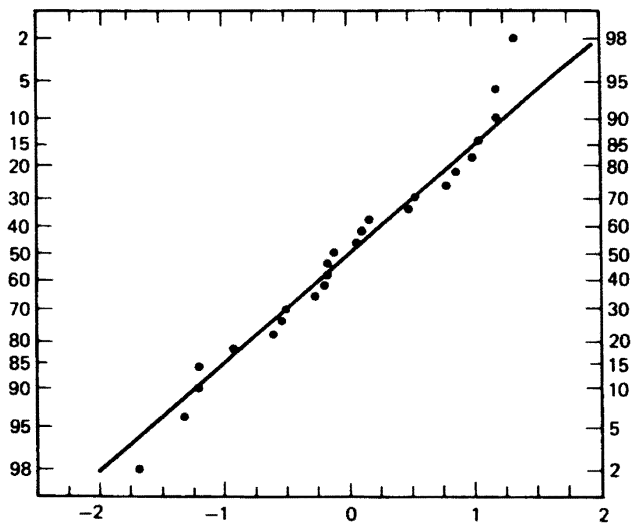


Figure 2A.4. Normal plot of the residuals of Table 1.2.

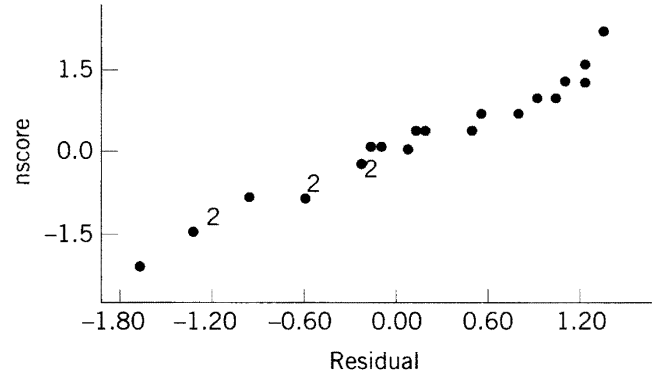


Figure 2A.5. Normal plot of the residuals of Table 1.2.

```
nscore c6  c7
plot c6  c7
```

would reverse the axes and change all the connected explanations. The point is a trivial one, but the consequences can be enormous. So watch it!

Example 1. Consider the $m = 25$ residuals given in Table 1.2. We first arrange these in ascending order, giving due regard to sign:

$$-1.68, -1.32, -1.20, -1.20, -0.93, -0.60, -0.53, -0.51, -0.26, -0.19, -0.17, -0.16, -0.12, 0.08, 0.11, 0.17, 0.50, 0.55, 0.80, 0.87, 1.00, 1.05, 1.20, 1.20, 1.34.$$

To obtain a full normal plot of these we set $m = 25$ in Eq. (2A.2) and successively set $i = 1, 2, \ldots, m$ to give the ordinate values:

$$2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62, 66, 70, 74, 78, 82, 86, 90, 94, 98.$$

These ordinate values are associated with the ordered residuals. Thus the bottom point in Figure 2A.4 is at abscissa -1.68, ordinate $= 2$, that is, $(-1.68, 2)$. Readers will note that to plot this, it is necessary to use the bottom and *right-hand* scale on the probability paper. The left-hand scale shows this point as $(-1.68, 98)$, the left-hand scale being (100 − right-hand scale); this is a peculiarity of probability paper that seems destined to persist. The second point plotted is at $(-1.32, 6)$ and so on. The line shown is drawn by eye and represents an attempted rough fit to the majority of the points with somewhat more weight given to the central points. Usually the

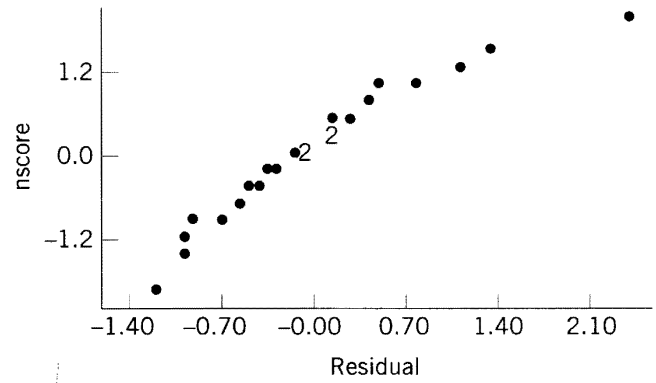
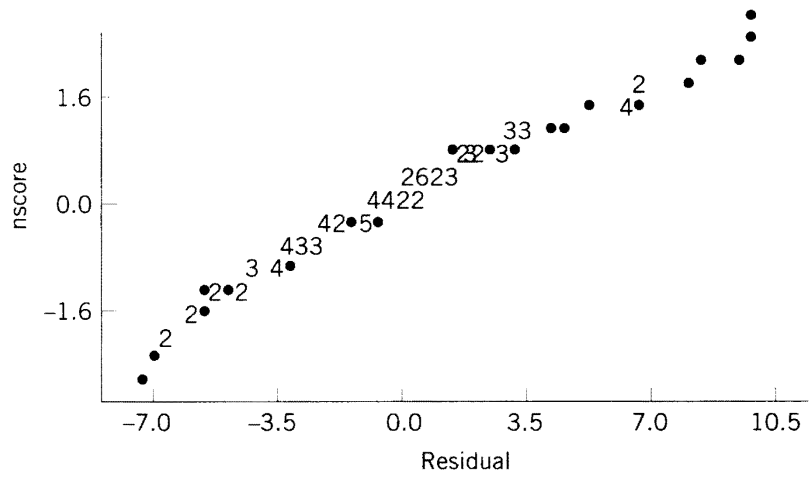
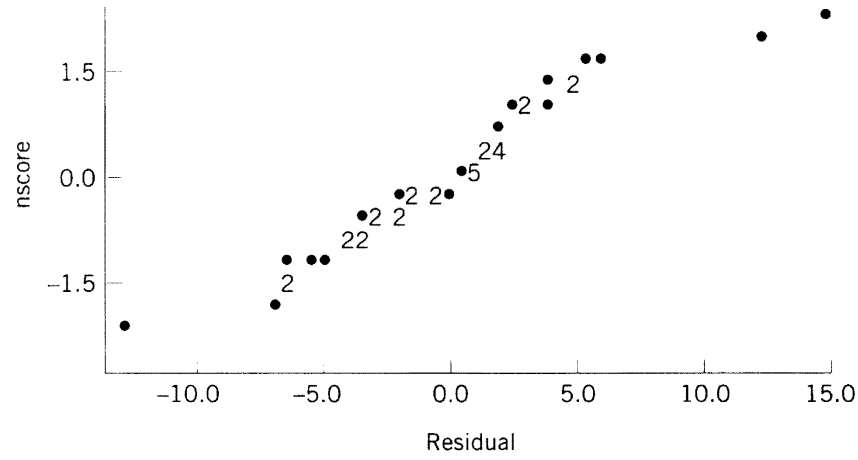


Figure 2A.6. Normal plot of the residuals from Section 2.1.



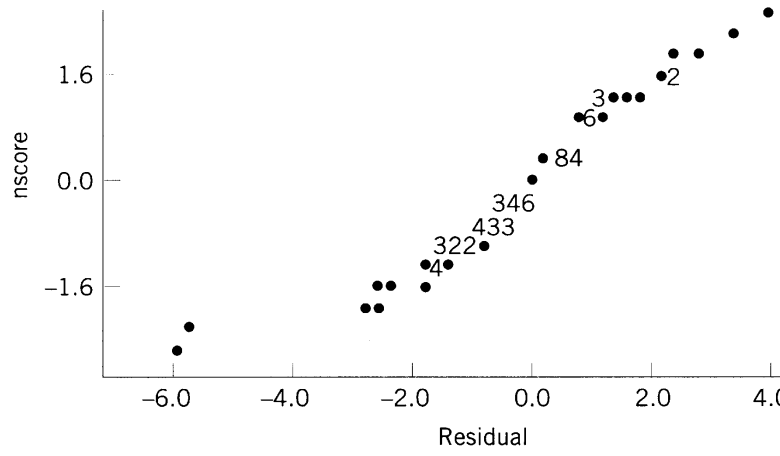
(*a*) Normal sample exhibiting lighter-tail characteristics



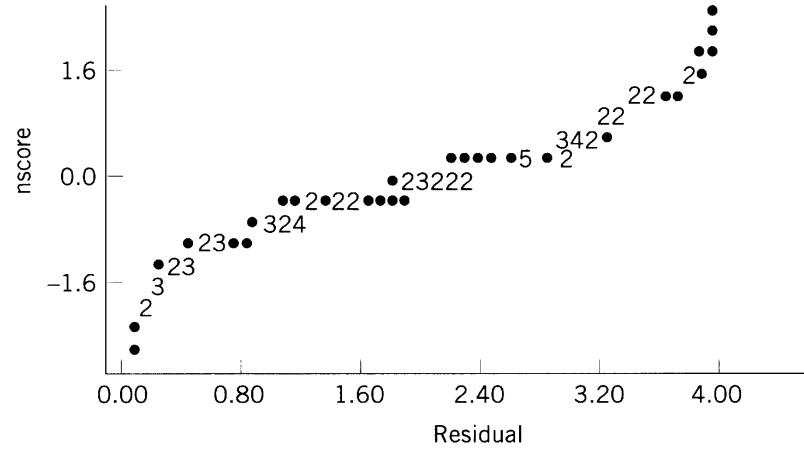
(*b*) Sample with one negative and two positive outliers

Figure 2A.7. Some characteristics of normal plots.

abscissa at which the line cuts the 50 ordinate would provide an estimate of the mean of the sample plotted but, in fact, we have drawn the line through (0, 50) here because the residuals sum to zero in theory. (In practice, rounding errors may occur, as we have noted.) An estimate of the standard deviation is $x_{84.13} - x_{50} = 0.97 - 0 = 0.97$ approximately. This compares well with $s = (0.7923)^{1/2} = 0.89$ from Table 1.6. The normal plot is not atypical of plots of normal samples of this size. The lowest two values and the highest two values are both "pulled in," to a minor extent, but this is not uncommon with least squares residuals. Certainly there are no outliers, which would show up as points well out to the left of the lower part of the line and well out to the right of the upper part of the line.



(*c*) Distribution "heavier-tailed" than the normal (e.g., *t*)



(*d*) Distribution "lighter-tailed" than the normal (e.g., uniform)

Figure 2A.7. (*Continued*)

Example 2. Figure 2A.5 shows a MINITAB computed plot of residuals versus nscores for the steam data. The plot is essentially identical to that of Figure 2A.4 except that the nscores were derived from MINITAB's $100(i - \frac{3}{8})/(m + \frac{1}{4})$ rather than our suggested $100(i - \frac{1}{2})/m$.

Example 3. Figure 2A.6 shows a MINITAB constructed plot of the residuals from the pure error example of Section 2.1. The outlying observation previously remarked upon appears off and to the right. It is too large to fall nicely on a line through the central bulk of the points.

Outliers

An outlier among residuals is one that is far greater than the rest in absolute value and perhaps lies three or four standard deviations or further from the mean of the residuals. The outlier is a peculiarity and indicates a data point that is not at all typical

of the rest of the data. It follows that an outlier should be submitted to particularly careful examination to see if the reason for its peculiarity can be determined.

Rules have been proposed for rejecting outliers [i.e., for deciding to remove the corresponding observation(s) from the data, after which the data are reanalyzed without these observations]. Automatic rejection of outliers is not always a very wise procedure. Sometimes the outlier is providing information that other data points cannot due to the fact that it arises from an unusual combination of circumstances, which may be of vital interest and requires further investigation rather than rejection. As a general rule, outliers should be rejected out of hand only if they can be traced to causes such as errors in recording the observations or in setting up the apparatus. Otherwise careful investigation is in order.

Some General Characteristics of Normal Plots

The diagrams in Figure 2A.7 show characteristics that can occur when residuals are displayed in a probability plot. In all diagrams, the probability or normal score is on the vertical axis and the residuals values are on the horizontal axis.

Normal plots are also used to examine effects (contrasts) from factorial experiments. In that context, the plot sometimes exhibits the look of two parallel lines. This typically indicates that (at least) one of the observations is suspect. See Box and Draper (1987, p. 132).

Making Your Own Probability Paper

The books of our childhood often had "projects for a rainy afternoon." Here is such a project. Probability paper on which the cumulative distribution curve becomes a straight line can be constructed for *any* continuous distribution as follows. Draw the cumulative distribution function. Draw horizontal lines at equal intervals of the vertical probability scale 0 to 1. At the points where the horizontal lines hit the curve, drop perpendiculars onto any horizontal line l, labeling the foot of the perpendicular 100 times the vertical probability scale reading from which it arose. The scale on the horizontal line l then provides the new spacings that should be employed on the vertical scale of the probability paper. Effectively, we have applied the inverse transformation $x = F^{-1}(y)$, where $y = F(x)$ is the cumulative probability function, to equal intervals of y. In labeling the new vertical axis, we multiply by 100 for convenience.

APPENDIX 2B. MINITAB INSTRUCTIONS

The MINITAB program below will obtain many of the details discussed in Chapters 1 and 2.

```
oh=0
set c2
2.3 1.8 2.8 1.5 2.2 3.8 1.8 3.7 1.7 2.8 &
2.8 2.2 3.2 1.9 1.8 3.5 2.8 2.1 3.4 3.2 &
3 3 5.9
set c1
1.3 1.3 2 2 2.7 3.3 3.3 3.7 3.7 4 &
4 4 4.7 4.7 5 5.3 5.3 5.3 5.7 6 &
6 6.3 6.7
set c20
12 23 7 8 17 22 1 11 19 20 &
```

```
5 2 21 15 18 3 6 10 4 9 13 14 16
end of data
corr c1 c2 m1

regress c2 1 c1 c11 c12;
resi c14;
pure.

plot c1 c2
print m1
print c11 c12
histogram c14
plot c14 c12
plot c14 c1
plot c14 c20
nscore c14 c41
plot c41 c14
end
stop
-------------------------
Comments:
The data are from Table 2.1.
c2 contains the Y's, c1 the X's.
c20 contains the time order.
c11 contains internally studentized residuals (Section 8.1).
c12 contains fitted values.
c14 contains residuals.
c41 contains nscores for the residual.
```

EXERCISES

Exercises for Chapter 2 are located in the section "Exercises for Chapters 1–3" at the end of Chapter 3.