

11-7.1 RESIDUAL ANALYSIS

The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$ where y_i is an actual observation and \hat{y}_i is the corresponding fitted value from the regression model. Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance and in determining whether additional terms in the model would be useful.

As an approximate check of normality, the experimenter can construct a frequency histogram of the residuals or a **normal probability plot of residuals**. Many computer programs will produce a normal probability plot of residuals, and because the sample sizes in regression are often too small for a histogram to be meaningful, the normal probability plotting method is preferred. It requires judgment to assess the abnormality of such plots. (Refer to the discussion of the “fat pencil” method in Section 6-6).

We may also **standardize** the residuals by computing $d_i = e_i / \sqrt{\hat{\sigma}^2}$, $i = 1, 2, \dots, n$. If the errors are normally distributed, approximately 95% of the standardized residuals should fall in the interval $(-2, +2)$. Residuals that are far outside this interval may indicate the presence of an **outlier**, that is, an observation that is not typical of the rest of the data. Various rules have been proposed for discarding outliers. However, they sometimes provide important information about unusual circumstances of interest to experimenters and should not be automatically discarded. For further discussion of outliers, see Montgomery, Peck, and Vining (2012).

It is frequently helpful to plot the residuals (1) in time sequence (if known), (2) against the \hat{y}_i , and (3) against the independent variable x . These graphs will usually look like one of the four general patterns shown in Fig. 11-9. Pattern (a) in Fig. 11-9 represents the ideal situation, and patterns (b), (c), and (d) represent anomalies. If the residuals appear as in (b), the variance of the observations may be increasing with time or with the magnitude of y_i or x_i . Data transformation on the response y is often used to eliminate this problem. Widely used variance-stabilizing transformations include the use of \sqrt{y} , $\ln y$, or $1/y$ as the response. See

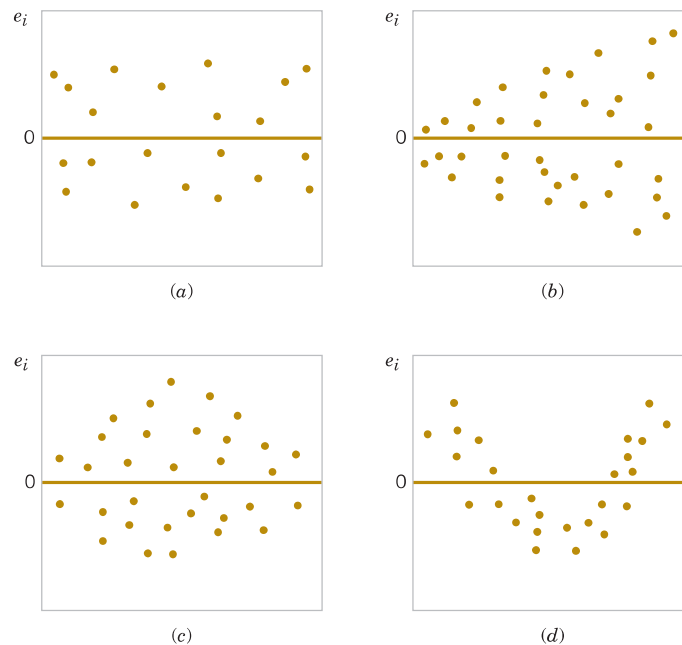


FIGURE 11-9 Patterns for residual plots. (a) Satisfactory. (b) Funnel. (c) Double bow. (d) Nonlinear. [Adapted from Montgomery, Peck, and Vining (2012).]

Montgomery, Peck, and Vining (2012) for more details regarding methods for selecting an appropriate transformation. Plots of residuals against \hat{y}_i and x_i that look like (c) also indicate inequality of variance. Residual plots that look like (d) indicate model inadequacy; that is, higher order terms should be added to the model, a transformation on the x -variable or the y -variable (or both) should be considered, or other regressors should be considered.

Example 11-7

Oxygen Purity Residuals The regression model for the oxygen purity data in Example 11-1 is $\hat{y} = 74.283 + 14.947x$. Table 11-4 presents the observed and predicted values of y at each value of x from this data set along with the corresponding residual. These values were calculated using a computer and show the number of decimal places typical of computer output.

A normal probability plot of the residuals is shown in Fig. 11-10. Because the residuals fall approximately along a straight line in the figure, we conclude that there is no severe departure from normality. The residuals are also plotted against the predicted value \hat{y}_i in Fig. 11-11 and against the hydrocarbon levels x_i in Fig. 11-12. These plots do not indicate any serious model inadequacies.

11-7.2 COEFFICIENT OF DETERMINATION (R^2)

A widely used measure for a regression model is the following ratio of sum of squares.

 R^2







The **coefficient of determination** is

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad (11-34)$$

The coefficient is often used to judge the adequacy of a regression model. Subsequently, we will see that in the case in which X and Y are jointly distributed random variables, R^2 is the square of the correlation coefficient between X and Y . From the analysis of variance identity in Equations 11-24 and 11-25, $0 \leq R^2 \leq 1$. We often refer loosely to R^2 as the amount of variability in the data explained or accounted for by the regression model. For the oxygen purity regression model, we have $R^2 = SS_R / SS_T = 152.13 / 173.38 = 0.877$; that is, the model accounts for 87.7% of the variability in the data.

TABLE • 11-4 Oxygen Purity Data from Example 11-1, Predicted \hat{y} Values, and Residuals

	Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, \hat{y}	Residual $e = y - \hat{y}$		Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, \hat{y}	Residual $e = y - \hat{y}$
1	0.99	90.01	89.081	0.929	11	1.19	93.54	92.071	1.469
2	1.02	89.05	89.530	-0.480	12	1.15	92.52	91.473	1.047
3	1.15	91.43	91.473	-0.043	13	0.98	90.56	88.932	1.628
4	1.29	93.74	93.566	0.174	14	1.01	89.54	89.380	0.160
5	1.46	96.73	96.107	0.623	15	1.11	89.85	90.875	-1.025
6	1.36	94.45	94.612	-0.162	16	1.20	90.39	92.220	-1.830
7	0.87	87.59	87.288	0.302	17	1.26	93.25	93.117	0.133
8	1.23	91.77	92.669	-0.899	18	1.32	93.41	94.014	-0.604
9	1.55	99.42	97.452	1.968	19	1.43	94.98	95.658	-0.678
10	1.40	93.65	95.210	-1.560	20	0.95	87.33	88.483	-1.153

- 12-61.**  **Go Tutorial** Consider the regression model fit to the coal and limestone mixture data in Exercise 12-17. Use density as the response.
- Calculate 90% confidence intervals on each regression coefficient.
 - Calculate a 90% confidence interval on mean density when the dielectric constant = 2.3 and the loss factor = 0.025.
 - Calculate a prediction interval on density for the same values of the regressors used in part (b).
- 12-62.**  **+** Consider the regression model fit to the nisin extraction data in Exercise 12-18.
- Calculate 95% confidence intervals on each regression coefficient.
 - Calculate a 95% confidence interval on mean nisin extraction when $x_1 = 15.5$ and $x_2 = 16$.
 - Calculate a prediction interval on nisin extraction for the same values of the regressors used in part (b).
 - Comment on the effect of a small sample size to the widths of these intervals.
- 12-63.** Consider the regression model fit to the gray range modulation data in Exercise 12-19. Use the useful range as the response.
- Calculate 99% confidence intervals on each regression coefficient.
 - Calculate a 99% confidence interval on mean useful range when brightness = 70 and contrast = 80.
 - Calculate a prediction interval on useful range for the same values of the regressors used in part (b).
 - Calculate a 99% confidence interval and a 99% a prediction interval on useful range when brightness = 50 and contrast = 25. Compare the widths of these intervals to those calculated in parts (b) and (c). Explain any differences in widths.
- 12-64.**  Consider the stack loss data in Exercise 12-20.
- Calculate 95% confidence intervals on each regression coefficient.
 - Calculate a 95% confidence interval on mean stack loss when $x_1 = 80$, $x_2 = 25$ and $x_3 = 90$.
 - Calculate a prediction interval on stack loss for the same values of the regressors used in part (b).
 - Calculate a 95% confidence interval and a 95% prediction interval on stack loss when $x_1 = 80$, $x_2 = 19$, and $x_3 = 93$. Compare the widths of these intervals to those calculated in parts (b) and (c). Explain any differences in widths.
- 12-65.**  **+** Consider the NFL data in Exercise 12-21.
- Find 95% confidence intervals on the regression coefficients.
 - What is the estimated standard error of $\hat{\mu}_{\tau_{K_0}}$ when the percentage of completions is 60%, the percentage of TDs is 4%, and the percentage of interceptions is 3%.
 - Find a 95% confidence interval on the mean rating when the percentage of completions is 60%, the percentage of TDs is 4%, and the percentage of interceptions is 3%.
- 12-66.**  **+** Consider the heat-treating data from Exercise 12-14.
- Find 95% confidence intervals on the regression coefficients.
 - Find a 95% confidence interval on mean PITCH when TEMP = 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.
 - Fit a model to PITCH using regressors $x_1 = \text{SOAKTIME} \times \text{SOAKPCT}$ and $x_2 = \text{DIFFTIME} \times \text{DIFFPCT}$. Using the model with regressors x_1 and x_2 , find a 95% confidence interval on mean PITCH when SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.
 - Compare the length of this confidence interval with the length of the confidence interval on mean PITCH at the same point from part (b), which used an additive model in SOAKTIME, SOAKPCT, DIFFTIME, and DIFFPCT. Which confidence interval is shorter? Does this tell you anything about which model is preferable?
- 12-67.**  **+** Consider the gasoline mileage data in Exercise 12-11.
- Find 99% confidence intervals on the regression coefficients.
 - Find a 99% confidence interval on the mean of Y for the regressor values in the first row of data.
 - Fit a new regression model to these data using *cid*, *etw*, and *axle* as the regressors. Find 99% confidence intervals on the regression coefficients in this new model.
 - Compare the lengths of the confidence intervals from part (c) with those found in part (a). Which intervals are longer? Does this offer any insight about which model is preferable?
- 12-68.** Consider the NHL data in Exercise 12-22.
- Find a 95% confidence interval on the regression coefficient for the variable *GF*.
 - Fit a simple linear regression model relating the response variable to the regressor *GF*.
 - Find a 95% confidence interval on the slope for the simple linear regression model from part (b).
 - Compare the lengths of the two confidence intervals computed in parts (a) and (c). Which interval is shorter? Does this tell you anything about which model is preferable?

12-5 Model Adequacy Checking

12-5.1 RESIDUAL ANALYSIS

The **residuals** from the multiple regression model, defined by $e_i = y_i - \hat{y}_i$, play an important role in judging model adequacy just as they do in simple linear regression. As noted in Section 11-7.1, several residual plots are often useful; these are illustrated in Example 12-10. It is also helpful to plot the residuals against variables not presently in the model that are possible candidates for inclusion. Patterns in these plots may indicate that the model may be improved by adding the candidate variable.

Example 12-10 Wire Bond Strength Residuals The residuals for the model from Example 12-1 are shown in Table 12-3. A normal probability plot of these residuals is shown in Fig. 12-6. No severe deviations from normality are obviously apparent, although the two largest residuals ($e_{15} = 5.84$ and $e_{17} = 4.33$) do not fall extremely close to a straight line drawn through the remaining residuals.

The standardized residuals

Standardized Residual

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\sqrt{\hat{\sigma}^2}} \quad (12-42)$$

are often more useful than the ordinary residuals when assessing residual magnitude. For the wire bond strength example, the standardized residuals corresponding to e_{15} and e_{17} are $d_{15} = 5.84 / \sqrt{5.2352} = 2.55$ and $d_{17} = 4.33 / \sqrt{5.2352} = 1.89$, and they do not seem unusually large. Inspection of the data does not reveal any error in collecting observations 15 and 17, nor does it produce any other reason to discard or modify these two points.

The residuals are plotted against \hat{y} in Fig. 12-7, and against x_1 and x_2 in Figs. 12-8 and 12-9, respectively.* The two largest residuals, e_{15} and e_{17} , are apparent. Figure 12-8 gives some indication that the model underpredicts the pull strength for assemblies with short wire length ($x_1 \leq 6$) and long wire length ($x_1 \geq 15$) and overpredicts the strength for assemblies with intermediate wire length ($7 \leq x_1 \leq 14$). The same impression is obtained from Fig. 12-7. Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2 , say, be added to the model) or other regressor variables not presently in the model affected the response.

In the wire bond strength example, we used the standardized residuals $d_i = e_i / \sqrt{\hat{\sigma}^2}$ as a measure of residual magnitude. Some analysts prefer to plot standardized residuals instead of ordinary residuals because the standardized residuals are scaled so that their standard deviation is approximately unity. Consequently, large residuals (that may indicate possible outliers or unusual observations) will be more obvious from inspection of the residual plots.

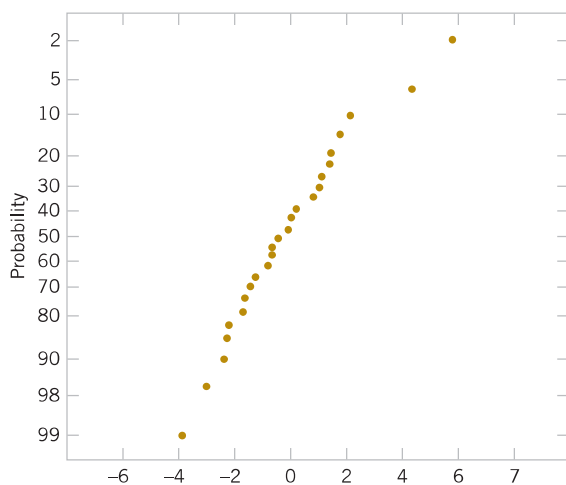


FIGURE 12-6 Normal probability plot of residuals.

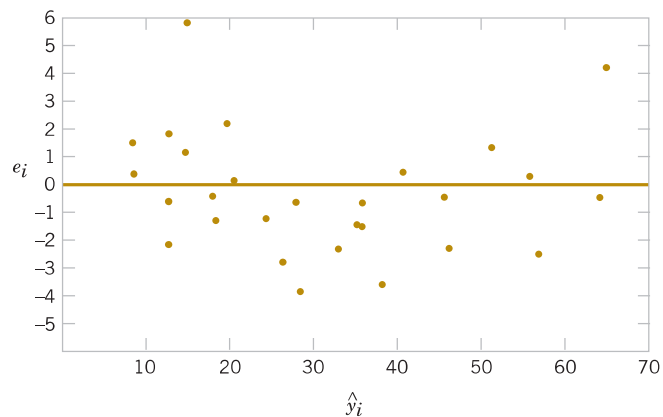
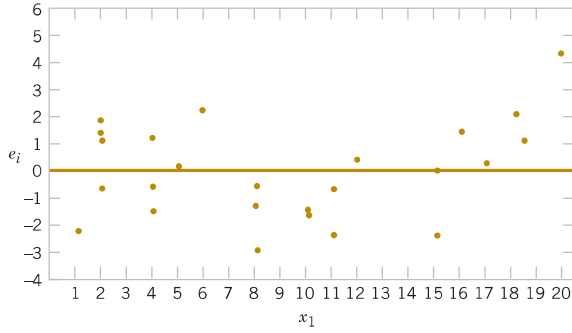
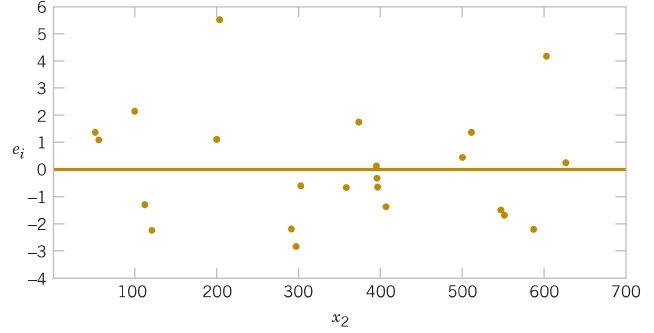


FIGURE 12-7 Plot of residuals against \hat{y} .

*There are other methods, such as those described in Montgomery, Peck, and Vining (2012) and Myers (1990), that plot a modified version of the residual, called a **partial residual**, against each regressor. These partial residual plots are useful in displaying the relationship between the response y and each individual regressor.

FIGURE 12-8 Plot of residuals against x_1 .FIGURE 12-9 Plot of residuals against x_2 .

Many regression computer programs compute other types of scaled residuals. One of the most popular are the **studentized residuals**

Studentized Residual

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \quad i = 1, 2, \dots, n \quad (12-43)$$

where h_{ii} is the i th diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The \mathbf{H} matrix is sometimes called the “**hat**” **matrix**, because

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Thus, \mathbf{H} transforms the observed values of \mathbf{y} into a vector of fitted values $\hat{\mathbf{y}}$.

Because each row of the matrix \mathbf{X} corresponds to a vector, say $\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$, another way to write the diagonal elements of the hat matrix is

Diagonal Elements of Hat Matrix

$$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad (12-44)$$

Note that apart from σ^2 , h_{ii} is the variance of the fitted value \hat{y}_i . The quantities h_{ii} were used in the computation of the confidence interval on the mean response in Section 12-3.2.

Under the usual assumptions that the model errors are independently distributed with mean zero and variance σ^2 , we can show that the variance of the i th residual e_i is

$$V(e_i) = \sigma^2(1 - h_{ii}), \quad i = 1, 2, \dots, n$$

Furthermore, the h_{ii} elements must fall in the interval $0 < h_{ii} \leq 1$. This implies that the standardized residuals understate the true residual magnitude; thus, the studentized residuals would be a better statistic to examine in evaluating potential **outliers**.

To illustrate, consider the two observations identified in the wire bond strength data (Example 12-10) as having residuals that might be unusually large, observations 15 and 17. The standardized residuals are

$$d_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2}} = \frac{5.84}{\sqrt{5.2352}} = 2.55 \quad \text{and} \quad d_{17} = \frac{e_{17}}{\sqrt{MS_E}} = \frac{4.33}{\sqrt{5.2352}} = 1.89$$