13 Nonlinear and Multiple Regression

13.1 Assessing Model Adequacy

Assessing Model Adequacy

A plot of the observed pairs (x_i, y_i) is a necessary first step in deciding on the form of a mathematical relationship between x and y.

It is possible to fit many functions other than a linear one ($y = b_0 + b_1x$) to the data, using either the principle of least squares or another fitting method.

Once a function of the chosen form has been fitted, it is important to check the fit of the model to see whether it is in fact appropriate.

Assessing Model Adequacy

One way to study the fit is to superimpose a graph of the best-fit function on the scatter plot of the data.

However, any tilt or curvature of the best-fit function may obscure some aspects of the fit that should be investigated.

Furthermore, the scale on the vertical axis may make it difficult to assess the extent to which observed values deviate from the best-fit function.

Residuals and Standardized Residuals

Residuals and Standardized Residuals

A more effective approach to assessment of model adequacy is to compute the fitted or predicted values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$ and then plot various functions of these computed quantities.

We then examine the plots either to confirm our choice of model or for indications that the model is not appropriate. Suppose the simple linear regression model is correct, and let $y = \hat{\beta}_0 + \hat{\beta}_1 x$ be the equation of the estimated regression line. Then the ith residual is $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

Residuals and Standardized Residuals

To derive properties of the residuals, let $e_i = Y_i - \hat{Y}_i$, represent the ith residual as a random variable (rv) before observations are actually made. Then

$$E(Y_i - \hat{Y}_i) = E(Y_i) - E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 x_i) = 0 \qquad \textbf{(13.1)}$$

Because $\hat{Y}_i (= \hat{\beta}_0 + \hat{\beta}_1 x_i)$ is a linear function of the Y_j's, so is $Y_i - \hat{Y}_i$ (the coefficients depend on the x_j's). Thus the normality of the Y_j's implies that each residual is normally distributed.

Residuals and Standardized Residuals

It can also be shown that

$$V(Y_i - \hat{Y}_i) = \sigma^2 \cdot \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \qquad \textbf{(13.2)}$$

Replacing σ^2 by s^2 and taking the square root of Equation (13.2) gives the estimated standard deviation of a residual.

Let's now standardize each residual by subtracting the mean value (zero) and then dividing by the estimated standard deviation.

Residuals and Standardized Residuals

The **standardized residuals** are given by

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}} \quad i = 1, \ldots, n \qquad \textbf{(13.3)}$$

If, for example, a particular standardized residual is 1.5, then the residual itself is 1.5 (estimated) standard deviations larger than what would be expected from fitting the correct model.

Residuals and Standardized Residuals

Notice that the variances of the residuals differ from one another. In fact, because there is a – sign in front of $(x_i - \bar{x})^2$, the variance of a residual decreases as x_i moves further away from the center of the data $\bar{x}$.

Intuitively, this is because the least squares line is pulled toward an observation whose x_i value lies far to the right or left of other observations in the sample.

Computation of e_i^*'s the can be tedious, but the most widely used statistical computer packages will provide these values and construct various plots involving them.

Example 1

We already know the presented data on x = burner area liberation rate and y = NO_x emissions.

Here we reproduce the data and give the fitted values, residuals, and standardized residuals. The estimated regression line is $y = -45.55 + 1.71x$, and $r^2 = .961$.

The standardized residuals are not a constant multiple of the residuals because the residual variances differ somewhat from one another.

Example 1

cont'd

x_i	y_i	$\hat{y}_i$	e_i	e_i^*
100	150	125.6	24.4	.75
125	140	168.4	−28.4	−.84
125	180	168.4	11.6	.35
150	210	211.1	−1.1	−.03
150	190	211.1	−21.1	−.62
200	320	296.7	23.3	.66
200	280	296.7	−16.7	−.47
250	400	382.3	17.7	.50
250	430	382.3	47.7	1.35
300	440	467.9	−27.9	−.80
300	390	467.9	−77.9	−2.24
350	600	553.4	46.6	1.39
400	610	639.0	−29.0	−.92
400	670	639.0	31.0	.99

Diagnostic Plots

13

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of model validity and usefulness are the following:

1. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis
2. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis
3. $\hat{y}_i$ on the vertical axis versus y_i on the horizontal axis
4. A normal probability plot of the standardized residuals

14

Diagnostic Plots

Plots 1 and 2 are called **residual plots** (against the independent variable and fitted values, respectively), whereas Plot 3 is fitted against observed values.

If Plot 3 yields points close to the 45° line [slope +1 through (0, 0)], then the estimated regression function gives accurate predictions of the values actually observed.

Thus Plot 3 provides a visual assessment of model effectiveness in making predictions. Provided that the model is correct, neither residual plot should exhibit distinct patterns.

15

Diagnostic Plots

The residuals should be randomly distributed about 0 according to a normal distribution, so all but a very few standardized residuals should lie between –2 and +2 (i.e., all but a few residuals within 2 standard deviations of their expected value 0).

The plot of standardized residuals versus $\hat{y}$ is really a combination of the two other plots, showing implicitly both how residuals vary with x and how fitted values compare with observed values.

16

Diagnostic Plots

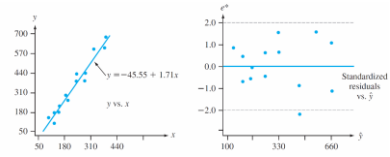
This latter plot is the single one most often recommended for multiple regression analysis.

Plot 4 allows the analyst to assess the plausibility of the assumption that ϵ has a normal distribution.

Example 2

Example 1. . . continued

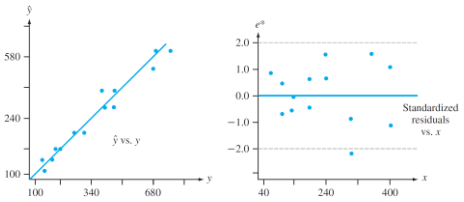
Figure 13.1 presents a scatter plot of the data and the four plots just recommended.



Plots for the data from Example 1

Figure 13.1

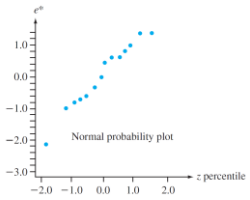
Example 2 cont'd



Plots for the data from Example 1

Figure 13.1

Example 2 cont'd



Plots for the data from Example 1

Figure 13.1

Example 2 cont'd

The plot of $\hat{y}$ versus y confirms the impression given by r^2 that x is effective in predicting y and also indicates that there is no observed y for which the predicted value is terribly far off the mark.

Both residual plots show no unusual pattern or discrepant values. There is one standardized residual slightly outside the interval (–2, 2), but this is not surprising in a sample of size 14.

The normal probability plot of the standardized residuals is reasonably straight. In summary, the plots leave us with no qualms about either the appropriateness of a simple linear relationship or the fit to the given data.

21

Difficulties and Remedies

22

Difficulties and Remedies

Although we hope that our analysis will yield plots like those of Figure 13.1, quite frequently the plots will suggest one or more of the following difficulties:

1. A nonlinear probabilistic relationship between x and y is appropriate.

2. The variance of ϵ (and of Y) is not a constant σ^2 but depends on x.

3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.

23

Difficulties and Remedies

4. The error term ϵ does not have a normal distribution.

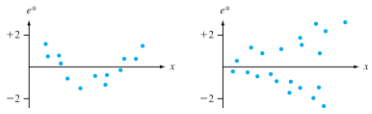
5. When the subscript i indicates the time order of the observations, the ϵ_i's exhibit dependence over time.

6. One or more relevant independent variables have been omitted from the model.

24

Difficulties and Remedies

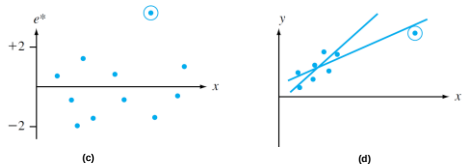
Figure 13.2 presents residual plots corresponding to items 1–3, 5, and 6. We discussed patterns in normal probability plots that cast doubt on the assumption of an underlying normal distribution.



Plots that indicate abnormality in data:
(a) nonlinear relationship; (b) nonconstant variance;

Figure 13.2

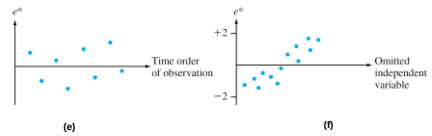
Difficulties and Remedies cont'd



Plots that indicate abnormality in data:
(c) discrepant observation; (d) observation with large influence;

Figure 13.2

Difficulties and Remedies cont'd



Plots that indicate abnormality in data:
(e) dependence in errors; (f) variable omitted

Figure 13.2

Difficulties and Remedies

Notice that the residuals from the data in Figure 13.2(d) with the circled point included would not by themselves necessarily suggest further analysis, yet when a new line is fit with that point deleted, the new line differs considerably from the original line.

This type of behavior is more difficult to identify in multiple regression. It is most likely to arise when there is a single (or very few) data point(s) with independent variable value(s) far removed from the remainder of the data.

We now indicate briefly what remedies are available for the types of difficulties.

Difficulties and Remedies

For a more comprehensive discussion, one or more of the references on regression analysis should be consulted. If the residual plot looks something like that of Figure 13.2(a), exhibiting a curved pattern, then a nonlinear function of x may be fit.

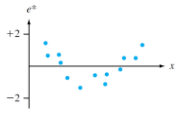


Figure 13.2(a)

Difficulties and Remedies

The residual plot of Figure 13.2(b) suggests that, although a straight-line relationship may be reasonable, the assumption that $V(Y_i) = \sigma^2$ for each i is of doubtful validity.

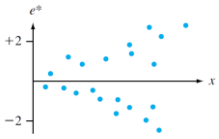


Figure 13.2(b)

Difficulties and Remedies

When the assumptions as studied earlier are valid, it can be shown that among all unbiased estimators of β_0 and β_1, the ordinary least squares estimators have minimum variance.

These estimators give equal weight to each (x_i, Y_i). If the variance of Y increases with x, then Y_i's for large x_i should be given less weight than those with small x_i. This suggests that β_0 and β_1 should be estimated by minimizing

$$f_w(b_0, b_1) = \Sigma w_i[y_i - (b_0 + b_1 x_i)]^2$$

where the w_i's are weights that decrease with increasing x_i.

Difficulties and Remedies

Minimization of Expression (13.4) yields **weighted least squares** estimates. For example, if the standard deviation of Y is proportional to x (for $x > 0$)—that is, $V(Y) = kx^2$—then it can be shown that the weights $w_i = 1/x_i^2$ yield best estimators of β_0 and β_1.

The books by John Neter et al. and by S. Chatterjee and Bertram Price contain more detail.

Weighted least squares is used quite frequently by econometricians (economists who use statistical methods) to estimate parameters.

Difficulties and Remedies

When plots or other evidence suggest that the data set contains outliers or points having large influence on the resulting fit, one possible approach is to omit these outlying points and recompute the estimated regression equation.

This would certainly be correct if it were found that the outliers resulted from errors in recording data values or experimental errors.

If no assignable cause can be found for the outliers, it is still desirable to report the estimated equation both with and without outliers omitted.

33

Difficulties and Remedies

Yet another approach is to retain possible outliers but to use an estimation principle that puts relatively less weight on outlying values than does the principle of least squares.

One such principle is MAD (minimize absolute deviations), which selects $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize $\Sigma |y_i - (b_0 + b_1 x_i|$.

Unlike the estimates of least squares, there are no nice formulas for the MAD estimates; their values must be found by using an iterative computational procedure.

34

Difficulties and Remedies

Such procedures are also used when it is suspected that the ϵ_i's have a distribution that is not normal but instead have "heavy tails" (making it much more likely than for the normal distribution that discrepant values will enter the sample); robust regression procedures are those that produce reliable estimates for a wide variety of underlying error distributions.

Least squares estimators are not robust in the same way that the sample mean $\overline{X}$ is not a robust estimator for μ.

35

Difficulties and Remedies

When a plot suggests time dependence in the error terms, an appropriate analysis may involve a transformation of the y's or else a model explicitly including a time variable.

Lastly, a plot such as that of Figure 13.2(f), which shows a pattern in the residuals when plotted against an omitted variable, suggests that a multiple regression model that includes the previously omitted variable should be considered.

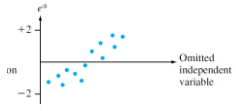


Figure 13.2(f)

36