The Normality Assumption

In the previous chapter we addressed problems with the least-squares fit that are primarily caused by individual data points or small subsets of data. In the current chapter we address larger problems where the data set as a whole fails to meet some model assumption. The text describes a common misunderstanding held by some researchers about least-squares procedures, namely the belief that the optimality results given by the Gauss-Markov Theorem when model assumptions hold imply optimal or even reasonable performance when model assumptions fail. Unfortunately this is a misunderstanding, and the truth is that least-squares procedures applied carelessly (as with any other statistical method applied carelessly) can lead to arbitrarily poor results. The conclusion that we should draw is that we should carefully check the data to assess model fit and to understand conclusions from the fitted model.

The first of these problems to discuss is the normality assumption. This assumption is checked by examining different graphical summaries of the residuals. A common display to check normality is a quantile-comparison plot. The procedure for creating this plot is described in Chapter 3 of the text, here is a partial summary assuming that we have the studentized residuals E_i^*:

1) Order the E_i^* values from smallest to largest, denoted by $E_{(1)}^*, E_{(2)}^*, \ldots, E_{(n)}^*$. The $E_{(i)}^*$ values are called the order statistics of the sample of E_i^* values.

2) Compute a measure related to the cumulative proportion of the values less than $E_{(i)}^*$:

$$P_i = \frac{i - \frac{1}{2}}{n}.$$

3) Now compute the quantile function (inverse of the CDF) to find the z_i (standard normal) value that corresponds to P_i:

$$z_i = P^{-1}(\frac{i - \frac{1}{2}}{n}).$$

4) Plot the z_i against the $E_{(i)}^*$ values, using the latter as the vertical coordinates.

If the residuals follow the normal assumption, then this plot should be approximately linear with intercept 0 and slope 1. As mentioned in the text in Chapters 3 and 12, the plot can be enhanced by including a line

that corresponds to a normal fit, and point-wise confidence limits can be calculated based on normal theory or a parametric bootstrap approach.

The quantile-comparison plot is particularly useful for examing the tail behavior of the residuals. Another plot can be used to examine the center of the distribution, a good choice is a histogram with a nonparametric density estimate superimposed. A nonparametric density estimate is essentially a smoothed version of the histogram, one method called the kernel density estimate is described in Chapter 3.

As noted in the text, different problems with the normality assumption can impede the interpretation of the least-squares fit and/or lead to a respecification of the model. In many cases a transformation can address problems with the normality assumption. We will take a closer look at choosing transformations later in the chapter, but in the example in the text it is illustrated that problems with positive skewness are generally addressed by applying a power transformation Y^p to Y, using a power $p < 1$.