Unusual and Influential Data

As we have seen, the linear model is a widely applicable and powerful tool for data analysis. However, the assumptions of the linear model are not always met in practice, and the least-squares fitting criterion can be drastically affected by a even just a single discrepant data point. In this and the next chapters we will learn about how to diagnose problems with the fit of a linear model, and ways to address these problems.

Outliers, Leverage, and Influential Points

An outlier in a linear model is a data point that has a large residual, meaning that the actual value y_i differs alot from its predicted value $\widehat{y}_i$. A leverage point is a data point whose covariate values are very different than those of most of the data. An influential data point is one that can lead to a change in some aspect of the model, such as the coefficient estimates, standard errors, etc. The text uses the Davis data and the one misrecorded data point to illustrate these concepts.

Hat Values

Recall that the hat matrix is defined by $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, where the name indicates that $\mathbf{H}$ 'puts the hat on $\mathbf{y}$' via $\widehat{\mathbf{y}}$= $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. We can easily see that:

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j,$$

which shows that $\widehat{y}_i$ is a weighted average of all y_j values, where the h_{ij} values are the weights. The diagonal elements satisfy $h_i \equiv h_{ii} = \sum_{j=1}^{n} h_{ij}^2$ due to the idempotence of $\mathbf{H}$, they summarize the influence of y_i on all of the fitted values. It can be shown that $1/n \leq h_i \leq 1$ (for models with intercepts), and since $\mathbf{H}$ is idempotent of rank $k+1$, we also have that the trace of $\mathbf{H}$ is $k+1$, so $\sum_{i=1}^{n} h_i = k+1$ and thus $\overline{h} = (k+1)/n$. It can be shown that

$$h_i = \frac{1}{n} + \frac{1}{n-1}(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}),$$

where $\mathbf{S}$ is the covariance matrix of the x variables. The quadratic form $(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})$ is the square of the Mahalanobis distance between $\mathbf{x}_i$ and $\overline{\mathbf{x}}$, it gives the covariance-weighted distance between the observation $\mathbf{x}_i$ and the vector mean $\overline{\mathbf{x}}$. Thus we can see that high leverage observations with an $\mathbf{x}_i$ vector far from the center of the data can play a dominant role in the fit of the model.

Residuals

We saw in the previous chapter that $V(\mathbf{e}) = \sigma_\varepsilon^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \sigma_\varepsilon^2(\mathbf{I}_n - \mathbf{H})$, so for an individual residual E_i we have $V(E_i) = \sigma_\varepsilon^2(1 - h_i)$. This means that the residuals E_i have different variances, so to account for unequal variance we could use:

$$E_i' = \frac{E_i}{S_E\sqrt{1-h_i}},$$

which is called a standardized residual. However, the numerator and denominator of E_i' are not independent, so a better idea is to estimate σ_ε^2 from all observations except the current one, yielding the studentized residual:

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}},$$

where $S_{E(-i)}$ is calculated from a model refit to the other $n-1$ observations, leaving out the *ith* observation. The text notes that this terminology is not consistent, so one must check documentation of software to be sure of which residuals are being computed by a particular program. An alternative way of defining studentized residuals is via a mean-shift outlier model,

$$Y_j = \alpha + \beta_1 X_{j1} + \beta_2 X_{j2} + \cdots + \beta_k X_{jk} + \gamma D_j + \varepsilon_j,$$

where D_j is a dummy variable for the *ith* observation. To test $H_0 : \gamma = 0$ the test statistic $t_0 = \widehat{\gamma}/SE(\widehat{\gamma})$ follows a t_{n-k-2} distribution under H_0, and is equal to the studentized residual E_i^*. It can be shown that E_i' and E_i^* are related via:

$$E_i^* = E_i'\sqrt{\frac{n-k-2}{n-k-1-E_i'^2}},$$

thus simplifying the calculation of studentized residuals. Also for large sample sizes, the term under the square root is close to 1 and the hat values are typically small so that $E_i^* \approx E_i' \approx E_i/S_E$.

Testing for Outliers

In checking for unusually large residuals E_i^*, we usually do not have particular observations in mind to check before analyzing the data, so we must make a correction to the usual t_{n-k-2} distribution of E_i^* to account for post hoc testing. A Bonferroni adjustment can be used for testing the largest absolute E_i^*, in which we adjust the observed p-value p' via $p = 2np'$, to account for the sample size n of observations and the two-tailed nature of the test. An alternative approach due to Anscombe allows the researcher the opportunity to pay a specified premium in efficiency to purchase a policy for outlier rejection.