1 More on residual analysis

The jackknife residuals follow a t distribution with $n-k-2$ degrees of freedom when model assumptions hold, so it is possible to use them to check for outliers. Also, the hat or leverage values h_i are used to assess the extremeness of an observation in the k dimensional space of the covariates $x_1, ..., x_k$. The h_i values satisfy $0 \leq h_i \leq 1$, and the average h_i value is $\overline{h} = (k+1)/n$. A rough rule of thumb is to consider an observation to have excessive leverage if $h_i > 2(k+1)/n$. Another measure of the influence of an observation is given by its Cook's distance: $d_i = \sum_{j=0}^{k} (\widehat{\beta}_j - \widehat{\beta}_{j(-i)})^2$, which measures how much the regression coefficients $\widehat{\beta}_j$ change when the observation is removed.

2 Collinearity

Collinearity is a problem that occurs due to correlation among the x variables. For two covariates x_1 and x_2, they are collinear if one is a linear function of the other, in which case their regression coefficients cannot be uniquely estimated. Generally, for k covariates $x_1, ..., x_k$, collinearity (also called multicollinearity) exists if some x variable is a linear function of the other x variables, such as $x_1 = \alpha_0 + \alpha_2 x_2 + ... + \alpha_k x_k$, in which case the regression coefficients are not uniquely defined. A more common problem is near (multi) collinearity, in which one or more x variables are almost perfectly predicted by the other x's. We can measure collinearity of x_j via the R^2 value from a regression model in which we predict x_j from the other x's: $x_j = \alpha_0 + \alpha_1 x_1 + ... + \alpha_{j-1} x_{j-1} + \alpha_{j+1} x_{j+1} + ... + \alpha_k x_k + E$. The text uses $R^2_{x_j \cdot x_1,...,x_{j-1},x_{j+1},...,x_k}$ for this, we can just use R_j^2 to denote this value. Two related measures of collinearity are the variance inflation factor, $\text{VIF}_j = 1/(1-R_j^2)$, and the tolerance, $\text{Tolerance}_j = 1 - R_j^2$. Collinearity has a direct effect on the variance of the estimates because $s^2_{\widehat{\beta}_j}$ is proportional to VIF_j, which directly effects the Partial F test for β_j that we considered previously. As a rough guide, collinearity is considered problematic when $\text{VIF}_j > 10$, or equivalently when either $\text{Tolerance}_j < .1$ or $R_j^2 > .9$. The VIF can be obtained in SAS by using the VIF option on the Model statement, and the Tolerance value is routinely included with SYSTAT regression output. Note that the effects of collinearity are easily seen in our sandwich data example, even though VIF is only 5.5, somewhat less than the rough guide of 10. Another way to measure collinearity is via a principal component analysis of the correlation matrix of the x variables, which we will not consider in this course.

3 Remedies

Problems that show up in the residual analysis can be addressed in several ways. Transforming the dependent variable y can sometimes resolve problems due to heterogeneous variance or non-normality of residuals. Weighted least-squares analysis can sometimes address the problem of heterogeneous variance when there is information about how the variances differ. If some observations are clearly influential, but not due to obvious data recording errors, then one option is to analyze the data both with and without the influential points and to compare the results to understand the impact that the influential points have on the analysis. If collinearity is present (especially in the case of polynomial regression) a possible solution is to center the data by using $x^*_{ij} = x_{ij} - \overline{x}_j$ instead of the original x_{ij} values. Another approach in the case of collinearity is to remove redundant variables.