Detecting Collinearity

We have encountered the issue of collinearity in previous lectures, particularly as it relates to two issues:

1) If the model matrix X is not of full rank (we have sought to avoid this situation), and

2) When studying the least-squares estimators we saw that the standard error of a slope estimate B_j is:

$$SE(B_j) = \frac{1}{\sqrt{1-R_j^2}} \frac{S_E}{\sqrt{\sum (x_{ij} - \overline{x}_j)^2}},$$

where R_j^2 is from the regression to predict X_j from all of the other X_i's. Part of this expression,

$$VIF(B_j) = \frac{1}{1-R_j^2}$$

is called the variance inflation factor, and tells us how correlation between this covariate and the other covariates has increased the variance of B_j. Various suggestions have been made for when to be concerned with large VIF values, with some sources using 10 as a cutoff. The text describes how large VIF values lead to problems that can be seen in added-variable plots and in confidence intervals for regression coefficients. The text also has some very useful figures to illustrate the effect of collinearity on the residual sum of squares for the regression model, showing that for extreme collinearity the residual sum of squares has a flat shape indicating a less-clearly defined least-squares estimate.

Principal Component Analysis

Variance inflation factors give a good description of the effect of other covariates on a single covariate, but they do not provide a summary of the dependence structure of the entire set of covariates. This can be accomplished by conducting a principal component analysis on the correlation matrix of the standardized regressors, $\mathbf{R}_{XX} = [1/(n-1)]\mathbf{Z}'_X\mathbf{Z}_X$. It can be shown that the variance inflation factors $VIF(B_j) = VIF_j$ can be expressed as a function of the eigenvalues L_l and the principal component coefficients A_{jl} by

$$VIF_j = \sum_{l=1}^{k} \frac{A_{jl}^2}{L_l}.$$

A global measure of collinearity is the condition number, $K = \sqrt{L_1/L_k}$, a value of 10 or more is problematic and indicates that $\mathbf{R}_{XX}$ is ill conditioned. Individual condition indexes $K_j = \sqrt{L_1/L_j}$ can also be used to identify the number of collinear relationships.

Generalized Variance Inflation

Fox and Monette (1992) have done work to extend the concept of variance inflation to sets of regressors, such as dummy variables and polynomial regressors, they have developed what is called the generalized variance inflation factor.