Assignment 7 STP 526, Fall 2009, S. Lohr

Due on Wednesday, 18 November 2009

1. Read Sections 8.3.1-8.3.2 and 4.5 of the text.

2. Use the stackloss data from HW 6 for this problem. Find the condition indices and the variance decomposition proportions for the centered data. Are there indications of multicollinearity in the data?

3. Ridge regression. Many methods have been proposed in the statistical literature for remedying the effects of multicollinearity. Some of these methods involve using a biased estimator $\tilde{\boldsymbol{\beta}}$ rather than the unbiased OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

   Let's assume the $n \times p$ matrix $\mathbf{X}$ has full rank. One consequence of multicollinearity is that $\mathbf{X}'\mathbf{X}$ is close to being singular. Thus, to increase stability, Hoerl and Kennard suggested
$$\tilde{\boldsymbol{\beta}}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$$
   for fixed $k \geq 0$.

   (a) Show that $\mathbf{X}'\mathbf{X} + k\mathbf{I}$ is nonsingular, so $\tilde{\boldsymbol{\beta}}_k$ is well defined. (Hint: use positive definiteness)

   (b) What is the bias $E[\tilde{\boldsymbol{\beta}}_k] - \boldsymbol{\beta}$? (Hint: see 3.9 on the matrix handout.)

   (c) Find the covariance matrix of $\tilde{\boldsymbol{\beta}}_k$.

   (d) Try out ridge regression with the stackloss data you used in HW 6. Find $\tilde{\boldsymbol{\beta}}_k$ for different values of k, and construct a ridge trace diagram; i.e., plot $\tilde{\beta}_{ki}$ versus k for each coefficient i. For what k do the coefficients appear to stabilize?

4. Consider the standard regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\mathbf{X}$ is $n \times k$ with i^{th} row $\mathbf{x}_i'$, rank$(\mathbf{X}) = k$, $E[\varepsilon] = \mathbf{0}$ and $\text{Cov}[\varepsilon] = \sigma^2\mathbf{I}$. We often use externally studentized residuals (you know these from SAS as Rstudent),
$$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{\text{MSE}_{(i)}(1 - p_{ii})}},$$
   where p_{ii} is the leverage (the i^{th} diagonal element of $\mathbf{P_X}$) and $\text{MSE}_{(i)} = \text{SSE}_{(i)}/(n - k - 1)$ is the mean squared error when the model is fit to all of the observations except observation i. In this exercise you will find how the externally studentized residuals may be calculated as functions of the SSE, leverage and raw residuals, without refitting the regression equation.

   (a) Let $\mathbf{X}_{(i)}$ be the $\mathbf{X}$ matrix with row i removed [$\mathbf{X}_{(i)}$ is $(n-1) \times k$]. Assume $\mathbf{X}_{(i)}$ has rank k for all i. Show that $\mathbf{X}_{(i)}'\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'$, and then use Result 1.2.10 to show
$$[\mathbf{X}_{(i)}'\mathbf{X}_{(i)}]^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}.$$

(b) Let $\mathbf{P}_{(i)}$ be the projection matrix onto $\mathcal{C}(\mathbf{X}_{(i)})$. Find an $(n-1)\times n$ matrix $\mathbf{A}$ of ones and zeroes such that $\mathbf{X}_{(i)} = \mathbf{A}\mathbf{X}$. Show that

$$\mathbf{P}_{(i)} = \mathbf{A}\mathbf{P}_{\mathbf{X}}\mathbf{A}' + \frac{\mathbf{A}\mathbf{P}_{\mathbf{X}}\mathbf{e}_i\mathbf{e}_i'\mathbf{P}_{\mathbf{X}}\mathbf{A}'}{1-p_{ii}}$$

where $\mathbf{e}_i$ is the vector with 1 in position i and zeroes elsewhere.

(c) Show, using the previous result, that

$$\mathrm{SSE}_{(i)} = \mathbf{y}_{(i)}'(\mathbf{I}-\mathbf{P}_{(i)})\mathbf{y}_{(i)} = \mathrm{SSE} - \frac{\hat{\varepsilon}_i^2}{1-p_{ii}}.$$

Hint: Write $\mathbf{A}'\mathbf{A} = \mathbf{I} - \mathbf{e}_i\mathbf{e}_i'$.

(d) Show that

$$r_i = \hat{\varepsilon}_i\sqrt{\frac{n-k-1}{\mathrm{MSE}\,(1-p_{ii})(n-k-\hat{\varepsilon}_i^2/[\mathrm{MSE}\,(1-p_{ii})])}}.$$

(e) Show that if ε is normally distributed, r_i^2 follows an F distribution with 1 and $n-k-1$ degrees of freedom. Use our matrix results to establish the necessary independence.