Advanced Methods in Biostatistics 1 (140.751)
Fall 2005

Homework Assignment #2
(Due Wednesday, October 5, 2005)

Please hand in a hard copy of your R code and send an electronic version of your R code to Kenny (kshum@jhsph.edu).

1. Let $\mathbf{G}$ be a generalized inverse of $\mathbf{X}'\mathbf{X}$. Show that
   (a) $\mathbf{G}'$ is also a generalized inverse of $\mathbf{X}'\mathbf{X}$,
   (b) $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}$, i. e. $\mathbf{G}\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$,
   (c) $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to $\mathbf{G}$,
   (d) $\mathbf{X}\mathbf{G}\mathbf{X}'$ is symmetric, whether $\mathbf{G}$ is or not.

   Hint for (b): Show first that $\mathbf{X}'\mathbf{X} = \mathbf{0}$ implies $\mathbf{X} = \mathbf{0}$, and second that $\mathbf{P}\mathbf{X}'\mathbf{X} = \mathbf{Q}\mathbf{X}'\mathbf{X}$ implies $\mathbf{P}\mathbf{X}' = \mathbf{Q}\mathbf{X}'$.

2. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Show that

$$\mathbf{b}^0 = \mathbf{G}\mathbf{X}'\mathbf{Y} + (\mathbf{G}\mathbf{X}'\mathbf{X} - \mathbf{I})\mathbf{z}$$

   is a solution to the normal equations for any vector $\mathbf{z}$ of the same length as $\boldsymbol{\beta}$, where $\mathbf{G}$ is any generalized inverse of $\mathbf{X}'\mathbf{X}$.

3. Assume that $\mathbf{P}_i$ is a projection matrix ($i = 1, 2$) and $\mathbf{P}_1 - \mathbf{P}_2$ is p.s.d. Show that
   (a) $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_2$ (Hint: first show that $\mathbf{P}_1\mathbf{x} = \mathbf{0}$ implies $\mathbf{P}_2\mathbf{x} = \mathbf{0}$).
   (b) $\mathbf{P}_1 - \mathbf{P}_2$ is a projection matrix.

4. Write an R function called `mylm()` that takes the response vector $\mathbf{Y}$ and the matrix of covariates $\mathbf{X}$ as input, and returns a list of the following:
   - **beta**, the vector of least squares estimates,
   - **sigma**, the residual standard error,
   - **varbeta**, the covariance matrix of the least squares estimates,
   - **fitted**, the vector of fitted values,
   - **residuals**, the vector of residuals.

   Further, put in an option to return **hat**, the projection matrix, upon request. The default should be to not return it. To fit an intercept, the elements in the first column of $\mathbf{X}$ have to be equal to one, so your function should also have an option to add a vector of ones to the matrix with the predictors. Further, your function should check whether or not $\mathbf{X}'\mathbf{X}$ is invertible, and stop if it is not. Find a dataset to try out your function (you can simulate one if you like), and compare the results to the one from the `lm()` function.