BST 760: Advanced Regression
Breheny

Assignment 3

Due: Thursday, February 3

1. Show that the inverse of a symmetric matrix is also symmetric.

2. Show that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent.

3. Show that second derivative (Hessian) of the residual sum of squares with respect to beta is equal to $2\mathbf{X}^T\mathbf{X}$.

4. Inverting large matrices can be quite computer intensive and can lead to instability if the matrix is close to singular. For these reasons, algorithms for linear regression (like the `lm` function in `R`, for example) do not actually invert $\mathbf{X}^T\mathbf{X}$ when they solve for the regression coefficients $\hat{\boldsymbol{\beta}}$. They rely on a shortcut called *QR decomposition*. For any full-rank $\mathbf{X}$, there exist matrices $\mathbf{Q}$ and $\mathbf{R}$ such that $\mathbf{X} = \mathbf{QR}$, $\mathbf{Q}$ is orthogonal ($\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is *upper triangular*. A matrix is upper triangular if all the entries below the main diagonal are 0. For example,

$$\begin{bmatrix} 3 & 2 & 1 & 2 \\ 0 & 2 & 5 & 3 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

   Note that a triangular matrix is not symmetric, but it is invertible.

   (a) Show that $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{Q}^T\mathbf{y}$.

   (b) Given the above equation, it is very easy to solve for $\hat{\boldsymbol{\beta}}$; why? (Hint: The technique is called *backsolving*, and it starts by solving for $\hat{\beta}_p$, then working backwards towards $\hat{\beta}_0$)

5. The psychologist Robert Levine has conducted a number of studies investigating the association between the "pace of life" and heart disease. The course web site contains the data from one such study, published in Levine, R.V. (1990). The Pace of Life, *American Scientist* 78: 450459. The data set contains measurements from 36 metropolitan areas throughout the U.S. on the following four variables:

   - `Heart`: Age-adjusted death rate due to heart disease
   - `Walk`: Average walking speed of downtown pedestrians
   - `Bank`: Average time taken by bank clerks to complete a standard request
   - `Talk`: Talking speed (syllables per second) of postal clerks

   All four variables are standardized and ordered so that high values in `Walk`, `Talk`, and `Bank` correspond to a high pace of life. Note: because all the variables are standardized, there is no need to include an intercept in this data set. Both `R` and SAS include an intercept by default. To suppress this, in `R` you can type a 0 when specifying the model terms, as in `Heart 0+Walk`, whereas in SAS, you can add a `NOINT` option in the model statement.

(a) Create a scatterplot matrix of the four variables in this data set.

(b) Fit a regression model with heart disease as the outcome variable and the three pace of life variables as explanatory variables. Report the coefficients you obtain.

(c) Describe the results of your regression model qualitatively. Do cities with a higher pace of life have higher death rates from heart disease?

(d) Talking speed is positively correlated with heart disease, but it has a negative regression coefficient in the above model. How is this possible?

(e) If walking speed goes up by 1 SD, but the other variables remain the same, how will that affect the average heart disease rate? Report a number $\pm$ a standard error.

(f) If both walking pace and bank pace go up by 1 SD, but talking pace remains the same, how will that affect the average heart disease rate? Report a number $\pm$ a standard error.

(g) If a city has an average talking pace and walking pace, but a banking pace two standard deviations above average, what will its heart disease rate be? Report a number $\pm$ a prediction error (*i.e.*, a number that captures the variability of both the model estimates and the random variability among cities).