FALL 2010
Stat 849: Homework Assignment 1
Due: September 24, 2010
Total points = 70

1. Suppose $\mathcal{Y}$ is $\mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ and suppose $\boldsymbol{X}$ is an $n \times p$ matrix of constants with rank $p < n$.

   (a) Show that $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{I} - \boldsymbol{A} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ are idempotent and find the rank of each.

   (b) if $\boldsymbol{\mu}$ is a linear combination of columns of $\boldsymbol{X}$, e.g., $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{b}$ for some $\boldsymbol{b}$, find $E(\mathcal{Y}'\boldsymbol{A}\mathcal{Y})$ and $E[\mathcal{Y}'(\boldsymbol{I} - \boldsymbol{A})\mathcal{Y}]$, where $\boldsymbol{A}$ is defined in part (a).

   (c) Find the distributions of $\mathcal{Y}'\boldsymbol{A}\mathcal{Y}/\sigma^2$ and $\mathcal{Y}'(I - \boldsymbol{A})\mathcal{Y}/\sigma^2$.

   (d) Show that $\mathcal{Y}'\boldsymbol{A}'\mathcal{Y}$ and $\mathcal{Y}'(I - \boldsymbol{A})\mathcal{Y}$ are independent.

   (e) Find the distribution of
   $$\frac{\mathcal{Y}'\boldsymbol{A}\mathcal{Y}/p}{\mathcal{Y}'(I - \boldsymbol{A})\mathcal{Y}/(n-p)}.$$

2. Consider the simple linear regression model
   $$Y = \beta_0 + \beta_1 X + \epsilon,$$
   where the intercept β_0 is known.

   (a) Find the least squares estimator of β_1 in this model.

   (b) Find variance of the estimator that you found in part (a). How does this compare with the least squares estimator of β_1 in a model where β_0 is not known?

3. Let
   $$\boldsymbol{A} = \begin{bmatrix} \frac{2}{3} & 0 & \frac{1}{3}\sqrt{2} \\ 0 & 1 & 0 \\ \frac{1}{3}\sqrt{2} & 0 & \frac{1}{3} \end{bmatrix}$$

   (a) Find the rank of $\boldsymbol{A}$.

   (b) Show that $\boldsymbol{A}$ is idempotent.

   (c) Show that $\boldsymbol{I} - \boldsymbol{A}$ is idempotent.

   (d) Show that $\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{A}) = \boldsymbol{0}$.

   (e) Find $\mathrm{tr}(\boldsymbol{A})$.

   (f) Find the eigenvalues of $\boldsymbol{A}$.

4. Set up the model matrix $\boldsymbol{X}$ and the $\boldsymbol{\beta}$ vector for each of the following regression models:

   (a) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} X_{i2} + \epsilon_i$, $i = 1, \ldots, 4$.

   (b) $\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$, $i = 1, \ldots, 4$.

   (c) $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$, $i = 1, \cdots, 5$.

   (d) $\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 + \log_{10} X_{i2} + \epsilon_i$, $i = 1, \cdots, 5$.

5. The data for exercise 1.19 in the book by Sen and Srivastava (`R` package `SenSrivastava`; data set E1.19) provide the price of books versus the number of pages and a characterization of whether the book is a paperback or a hardcover book.

   - Provide separate plots of price versus number pages by book type. Use the same axes for each plot.
   - Provide an overlaid plot of price versus number of pages using different symbols for the two types of books.
   - Which plot do you think is more effective and why?
   - Would you consider transforming the axes in these plots and, if so, how? Explain why or why not you would transform.
   - Provide a single "key graph" showing the relationship between the number of pages and the price on whatever scale you feel is suitable. The plot may be a multi-panel plot and may contain smoother lines. Provide a caption for your plot. Describe why you chose this plot and how this plot will influence your initial choice of a statistical model for these data.

   Hint: Here is how you can access these data in `R`:

```
> library(SenSrivastava)
> str(E1.19)
'data.frame':        20 obs. of  3 variables:
 $ Price: num  10.2 14.2 29.2 17.5 12 ...
 $ P    : num  112 260 250 382 175 146 212 292 340 252 ...
 $ B    : Factor w/ 2 levels "c","p": 2 2 1 2 2 1 1 1 2 1 ...
```

   Note that you must first install the `SenSrivastava` package using, for example

```
> install.packages("SenSrivastava")
```

6. A large, national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y). These data are available in the file `grocery_retailer.txt` on the course web site's data directory `http://www.stat.wisc.edu/~st849-1/data/`.

   You can read the data directly from the URL without needing to download

```
> str(groc <- read.table("http://www.stat.wisc.edu/~st849-1/data/grocery_retailer.txt",
+      header = TRUE))
'data.frame':        52 obs. of  4 variables:
 $ Y : int  4264 4496 4317 4292 4945 4325 4110 4111 4161 4560 ...
 $ X1: int  305657 328476 317164 366745 265518 301995 269334 26..
 $ X2: num  7.17 6.2 4.61 7.02 8.61 6.88 7.23 6.27 6.49 6.37 ...
 $ X3: int  0 0 0 0 1 0 0 0 0 0 ...
```

   (a) Provide various useful plots of these data (scatter plots etc...). What information can you gather from these plots?

   (b) Fit a linear regression model to these data. What are the estimated coefficients and standard errors of these estimates? How is the coefficient in front of holiday is interpreted?

(c) Investigate the residual plots. How well are the Gauss-Markov assumptions satisfied? Comment on anything unusual you see.

7. A scale has two pans. The measurements given by the scale is the difference between the weights in pan # 1 and pan # 2 plus a random error. Thus, if a weight μ_1 is put in pan # 1, a weight μ_2 is put in pan # 2, then the measurement is $Y = \mu_1 - \mu_2 + \epsilon$. Suppose that $E[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2$, and that in repeated uses of the scale, observations Y_i are independent.

   Suppose that two objects, #1 and #2, have weights β_1 and β_2. Measurements are taken as follows:

   (a) Object #1 is put on pan #1, nothing on pan # 2.

   (b) Object #2 is put on pan # 2, nothing on pan # 1.

   (c) Object #1 is put on pan # 1, object #2 on pan #2.

   (d) Objects #1 and #2 both put on pan # 1.

   Answer the following questions based on the above measurements.

   (a) Let $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4)'$ be the vector of observations. Formulate this as a linear model.

   (b) Find vectors $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ such that $\hat{\beta}_1 = \boldsymbol{c}_1'\boldsymbol{Y}$ and $\hat{\beta}_2 = \boldsymbol{c}_2'\boldsymbol{Y}$.

   (c) Find the covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$.