Stat 601 Homework 2 Part 2 Due 10/02/2015

1. Consider the simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon;$$

   (a) Find the least squares estimator of β_1 in this model assuming β_0 is known.

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

2. Consider the following three models:

   (1) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} X_{i2} + \epsilon_i,\ i = 1, \dots, 30.$

   (2) $1/Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,\ i = 1, \dots, 30.$

   (3) $Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i,\ i = 1, \dots, 30.$

   and a data set HWK2_Ext2.txt on Learn@UW. The goal is to estimate parameters $\beta_i, i = 0, 1, 2, 3$. The data contains a matrix of 3150 rows and 3 columns. You should use the link in the email and the following formula to choose a subset of the data to get estimates for the parameters.

   - Suppose your choice of data set is kth set.
   - The starting row number of your data subset is $(k-1) * 90 + 1$.
   - The ending row number of your data subset is $k * 90$.

   Now you get a data set with 90 rows and 3 columns. You need to divide the data into three subsets. The first 30 rows are for one of the three models; the middle 30 rows are for one of the three models; and the last 30 rows are for one of the three models. Within each subset, one column contains Y values; one column contains X_1; and one column contains X_2. They are not in any particular order. You will need to figure out which column is for Y, X_1 and X_2 respectively, and find the best fitted model.

3. 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")
```

4. 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.

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

```
> str(groc <- read.table("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.