Homework 9

(1) Suppose that

$$Y_i = \beta X_i + \epsilon_i.$$

(Recall that this is the regression-through-the-origin model.) Suppose that $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma_i^2$ where $\sigma_1^2, \ldots, \sigma_n^2$ are known constants. (As usual, treat the X_i's as constants.) For this question, do all your calculations directly. Do not use the general results about weighted regression in the lecture notes.

(a) Find the least squares estimator $\widehat{\beta}$.

(b) Find the weighted least squares estimator $\widetilde{\beta}$ based on minimizing $\sum_i (Y_i - \beta X_i)^2/\sigma_i^2$.

(c) Find the mean and variance of $\widehat{\beta}$ and $\widetilde{\beta}$.

(d) Show that $\text{Var}(\widetilde{\beta}) \leq \text{Var}(\widehat{\beta})$.

Hint: You may want to use the Cauchy-Schwarz inequality: $|\sum_i a_i b_i| \leq \sqrt{\sum_i a_i^2}\ \sqrt{\sum_i b_i^2}$.

(e) Suppose that $\epsilon_i \sim N(0, \sigma_i^2)$. Find the distribution of $\widetilde{\beta}$. Find a $1-\alpha$ confidence interval for β.

(f) To see the difference that weighting makes, do the following simulation. Generate data as follows:

```
n = 100
x = runif(n)
s = x^2
y = 3*x + rnorm(n,0,s)
```

Compute three estimators: the ordinary least squares estimator, the weighted least squares estimator (using $w_i = 1/s_i^2$) and using the weighted least squares estimator based on estimating the variances. (You can use the method in the example in Section 5.1 of Lecture notes 24).

Repeat this process 1,000 times. Report the variance of the three estimators. Plot a histogram of the estimators you get from each simulation. Make sure that the x-axis on each histogram has the same range, otherwise the histograms are hard to compare.

Based on your simulation, what is the better strategy if we don;t know the variances: (i) use ordinary least squares or (ii) use weighted least squares based on estimated variances.

(2) Get the Apple data.

```
library(alr4)
attach(allshoots)
names(allshoots)
help(allshoots)
```

The goal is to predict the number of 'stem units' on an apple tree based on days since dormancy. The variables are:

Day	days from dormancy
n	number of shoots sampled
ybar	average number of stem units
SD	within-day standard deviation
Type	1 if long shoots, 0 if shortshoots.

The data were collected on 106 days. But for each data, we do not have the raw data. Instead, we have the average value (ybar), the number of data points (n) and the standard deviation. Also, there are two types of shoots (this is the binary variable Type). The goal is to predict ybar from Days.

(a) Plot the data. Does the relationship between ybar and Days look different for the two different types?

(b) Perform a linear regression on ybar on Days and Type and their interaction. Summarize the fitted model. Check the residuals and report any problems you see (if any). Give 90 percent confidence intervals for the coefficients.

(c) If we assume a constant variance, then $\mathrm{Var}(\overline{y}_i) = \sigma^2/n_i \propto 1/n_i$. This suggests doing a weighted regression with weights based on the n_i's. Do the weighted regression. Give 90 percent confidence intervals for the coefficients and compare them to the confidence intervals from the unweighted regression.

(d) Using your fitted model from part(c), you now can construct a fitted line for the two types of shoots. Plot the data and add the two fitted lines. Comment on result.

(3) Download the following economic data:

```
library(alr4)
attach(BigMac2003)
names(BigMac2003)
help(BigMac2003)
```

The goal is to predict the variable FoodIndex (a measure of how expensive food is) from the other variables.

(a) Fit a linear regression model. Do all the usual diagnostics and comment on the diagnostics.

(b) There is a hypothesis that the price of a BigMac is a good predictor of FoodIndex. (The magazine *The Economist* has published a Big Mac index for many years.) Construct a 99 percent confidence interval for the BigMac variable in your regression. Interpret the confidence interval.

(c) To explore this further, fit a regression model that uses BigMac as the only covariate. Use an F-test to compare this to your previous model. What hypothesis are you testing? What is your conclusion?

(d) Now we want to address a different question: how good is each of your two models in terms of prediction error? Estimate the prediction error of both models. What is your conclusion?