FALL 2010
Stat 849: Homework Assignment 2
Due: October 8, 2010
Total points = 70

1. (50 pts) Consider the following multiple linear regression problem

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X_2 + \beta_3 X3 + \beta_4 X_4 + \epsilon.$$

Construct tests for the following hypotheses:

(a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$ vs. H_A : not so.

(b) $H_0 : \beta_1 = \beta_2, \beta_3 = \beta_4$ vs. H_A : not so.

(c) $H_0 : \beta_1 - 2\beta_2 = 4\beta_3, \quad \beta_1 + 2\beta_2 = 0$ vs. H_A : not so.

2. (50 pts) In this question, you are going to do a simulation study to compare the least squares estimator with the weighted least squares estimator. In particular, you are going to generate $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X\beta}, \boldsymbol{V})$, where $\boldsymbol{V}$ is a diagonal variance covariance matrix.

It happens that a simulation study of the results of `lm` fits can be done very effectively by creating a matrix of responses and using this as the response in the formula, see `http://www.stat.wisc.edu/~st849-1/Rnotes/Simulations2.html`

First we generate covariates randomly, assign the true parameter vector, and create the true mean response.

```
> ## n is the sample size in the regression model
> ## S is the number of simulations of the model fits
>
> ## Generate the covariates
> dat <- data.frame(X1 = rnorm(n, 3, 2), X2 = rexp(n, 1), X3 = rgamma(n, 2, 3))
> ## Assign the true coefficient vector
> betaTrue <- c(beta0 = 2, beta1 = 1.2, beta2 = -1.4, beta3 = -0.5)
> ## Calculate the true mean response
> muTrue <- as.vector(model.matrix(~ X1 + X2 + X3, dat) %*% betaTrue)
```

Next we generate the diagonal of $\boldsymbol{V}$, again using a random generator. It is a good idea to immediately convert this vector to standard deviations. You can use the `mvrnorm` function from the `MASS` package to generate multivariate normal noise terms from a general variance-covariance matrix, $\boldsymbol{V}$, but when $\boldsymbol{V}$ is diagonal that just becomes a complicated way of multiplying the individual elements by their standard deviations.

```
> SigmaSq <- runif(n, 0.5, 3)
> sigma <- sqrt(SigmaSq)
```

A single realization of the response is calculated as

```
> yy <- muTrue + sigma * rnorm(n)
```

To generate all `S` realizations of the response in the form of a matrix with n rows and S columns use

```
> YY <- muTrue + sigma * matrix(rnorm(n * S), nrow=n, ncol=S)
```

The remaining parts of the simulation, which you will need to fill in yourself, are

```
> ## Fit the simulated responses by ordinary least squares.
> ## Fit the responses by weighted least squares.
> ## Extract and store coefficient estimates from the ordinary least squares
> ## fit and the weighted least squares fit.
```

You can focus on all or one of the coefficients above (β_0, β_1, β_2, β_3) and you need to show that weighted least squares estimation in this scenario indeed provides better estimates of the coefficients. For example, you can do $10,000$ simulations and compute mean squared error of ordinary least squares and weighted least squares estimators based on these, e.g., $\frac{1}{S}\sum_{s=1}^{S}(\hat{\beta}_k^{OLS,s} - \beta_k)^2$, where S is the total number of simulations, $\hat{\beta}_k^{OLS,s}$ is the ordinary least squares estimates in the s-th simulation, and β_k is the *true* parameter value. You need to set S to a large number, e.g., 10,000.

Try different sample sizes and report other performance measures such as the bias and empirical standard errors of the estimated coefficients.

3. (50 pts) A study was conducted to investigate the relationship between the size of the ants and the distance at which they foraged. Ants were collected at various distances from the colony, weighed, and measured. Because an ant's weight provides a measure of how much food, or energy, it carries, and because headwidth measurements allow the ants to be classified by size, the data provides detailed information on the correlations among ant size, foraging distance, and energy supply. Some colonies develop "worker-conservative" foraging strategies, in which ants foraging at greater distances consume relatively more food: this minimizes the risk of starvation and leads to fewer deaths. Other colonies use strategies that conserve energy (the colonies overall supply of food). In this case long distance foragers, who are more likely to die, will consume less food so that their deaths will not be as much of a strain on the colony's food supply. The data were collected at the Sierra Nevada Aquatic Research Laboratory (SNARL) in the Great Basin Desert Province. Collection trays were placed into the ground at different distances from the entrance to the ant colonies' mounds, and any ants walking into them were trapped. Below is a brief description of the data collected and the data is available on the course web site.

   Colony: This is a number that identifies which colony the ant was taken from. A total of 10 colonies is considered.

   Distance: This indicates (in meters) how far from the mound's entrance the tray was replaced.

   Mass: Weight of the ant in milligrams. This variable is used as a measure of how much food (energy) each ant had.

   Headwith: A measure of the ant's maximum headwidth. Headwidth is a good indicator of an ant's size.

   You can access the data as

```
> DataURL <- "http://www.stat.wisc.edu/~st849-1/data/"
> ants <- read.table(paste(DataURL, "thatch_ant_c5del.txt", sep=''),
+                    header=TRUE)
> str(ants)
'data.frame':        1104 obs. of  6 variables:
 $ Colony       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Distance     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Mass         : int  109 120 94 61 72 134 94 113 111 106 ...
 $ Headwidth    : int  45 43 42 33 41 46 43 42 42 43 ...
 $ Headwidth.mm.: num  1.9 1.81 1.77 1.39 1.73 ...
 $ Class        : Factor w/ 5 levels "<30","30-34",..: 5 4 4 2 4 5 4 4 4 4 ...
```

(Note that it would make sense to convert `Colony` to a factor.)

(a) Provide a `pairs` or `lattice::splom` plot of the data. What are your initial observations regarding the relationship of `Headwidth` to `Colony` and `Distance`?

(b) Construct a linear regression model of `Headwidth` as a function of `Colony` and `Distance` and answer the following questions based on this linear model.

i. What are the dimensions of the design matrix? What is the row entry of the design matrix for an ant from colony 4 with a distance of 4?

ii. How do you interpret the coefficients in this model? Be explicit.

iii. Test whether `Headwidth` is related to `Colony` and `Distance`. Write down the hypothesiss being tested explicitly and report the test statistic, its distributions and the result of the test.

iv. Test whether `Headwidth` is related to Distance allowing the presence of `Colony` variable in the model. Write down the hypothesiss being tested explicitly and report the test statistic, its distributions and the result of the test.

(c) Scientists have reasons to believe that sizes of the ants from colonies 8 and 10 should be about half the size of the ants from other colonies. Sizes of the ants from the rest of the colonies are considered to be approximately equal. Test this hypothesis in the context of the above linear regression model.

(d) Scientists decide to include the `Mass` variable in the model based on the pair plot of the data. Extend the above regression model to include the `Mass` variable. Consider appropriate transformations if necessary.

(e) Consider the interaction plot given in Figure 1. Interpret this plot. Perform a test to

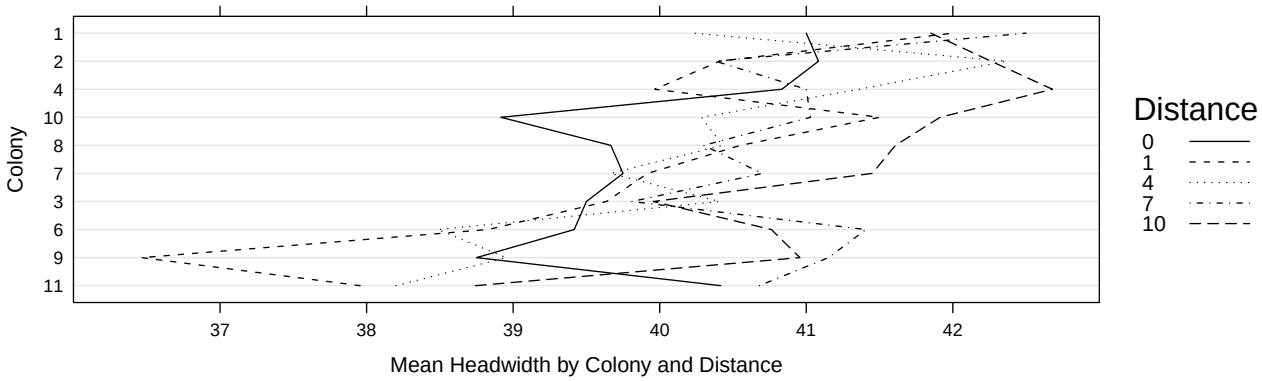

Figure 1: Interaction plot for part (e) of question 3

determine whether the data supports the general observation from this plot by restricting your full model NOT to include any interactions with the `Mass` variable. If you fail to reject the corresponding null hypothesis, can you think of a reason why that might be the case? Can you suggest another way of using the `Distance` variable in this analysis?

(f) If you can suggest another way of using the `Distance` variable in your analysis, implement this and compare the results with part (e).

4. (50 pts) Regulation of genes largely depend on the activities of site-specific DNA binding proteins called transcription factors. These factors recognize short DNA sequences in a specific manner, i.e, each nucleotide in a sequence defining a potential binding site is represented by one letter from the nucleotide alphabet $\{A, C, G, T\}$. The sequences at which these proteins bind are not unique. Binding can also occur at variants of the optimal site with some variants being more preferential for binding than others.

   In a recent study, `http://the_brain.bwh.harvard.edu/pubs/Bulyk02.pdf`, the DNA-binding specificities of Zif268 zinc fingers were investigated using microarray technology in an experiment that tested the feasibility of using microarrays for analyzing large number of DNA-protein interactions (if you are curious about what microarrays are, check out `http://www.bio.davidson.edu/Courses/genomics/chip/chip.html`). The data from this experiment contain binding measurements on all 64 possible 3mers, e.g., AAA, AAC, etc. For each 3mer, there are a total of 9 replicates and both the mean and standard error of data for each 3mer are reported. The `Ka` value of a sequence is proportional to the probability of Zif268 zinc finger protein binding to that sequence. You can access the data as

```
> dir <- "http://arep.med.harvard.edu/Bulyk/NAR2002supplementary/"
> str(Ka <- read.delim(paste(dir, "REDV_Ka_all9replicates.txt", sep = '')))
'data.frame':        64 obs. of  14 variables:
 $ X          : Factor w/ 64 levels "AAA","AAC","AAG",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ X.1        : logi  NA NA NA NA NA NA ...
 $ replic.1.Ka: num  0.000399 0.000399 0.000399 0.000399 0.000399 ...
 $ replic.2.Ka: num  0.000399 0.000399 0.000399 0.000399 0.000399 ...
 $ replic.3.Ka: num  0.000389 0.000389 0.000389 0.000389 0.000389 ...
 $ replic.4.Ka: num  0.000393 0.000393 0.000393 0.000393 0.000393 ...
 $ replic.5.Ka: num  0.000402 0.000402 0.000402 0.000402 0.000402 ...
 $ replic.6.Ka: num  0.000393 0.000393 0.000393 0.000393 0.000393 ...
 $ replic.7.Ka: num  0.000389 0.000389 0.000389 0.000389 0.000389 ...
 $ replic.8.Ka: num  0.000389 0.000389 0.000389 0.000389 0.000389 ...
 $ replic.9.Ka: num  0.000397 0.000397 0.000397 0.000397 0.000397 ...
 $ X.2        : logi  NA NA NA NA NA NA ...
 $ avg.Ka     : num  0.000394 0.000394 0.000394 0.000394 0.000394 ...
 $ SD.Ka      : num  5.03e-06 5.03e-06 5.03e-06 5.03e-06 5.03e-06 5.03e-06 5.79e-05 5.0
> Ka <- subset(Ka, select= -c(X.1, X.2))  # drop the extra columns
> names(Ka) <- c("trimer", paste("r", 1:9, sep=""), "avg", "sd")
```

   This form of the data is called the "wide" format. For analysis in R we want the "long" format where all the binding measurements are in one column. The `melt` function from the `reshape` package provides a convenient way of creating this.

```
> library(reshape)
> str(lKa <- melt(Ka[, 1:10], id = 1))
'data.frame':        576 obs. of  3 variables:
 $ trimer  : Factor w/ 64 levels "AAA","AAC","AAG",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ variable: Factor w/ 9 levels "r1","r2","r3",..: 1 1 1 1 1 1 1 1 1 ...
 $ value   : num  0.000399 0.000399 0.000399 0.000399 0.000399 ...
```

   For the purpose of modeling we want to regard the nucleotide in each of the three positions as a separate covariate.

```
> lKa <- within(lKa,
+           {
+               p3 <- factor(substring(trimer, 3, 3))
```

```
+               p2 <- factor(substring(trimer, 2, 2))
+               p1 <- factor(substring(trimer, 1, 1))
+           })
> names(lKa) <- c("trimer", "rep", "binding", "p1", "p2", "p3")
> head(lKa)
  trimer rep     binding p1 p2 p3
1    AAA  r1 0.000398571  A  A  A
2    AAC  r1 0.000398571  A  A  C
3    AAG  r1 0.000398571  A  A  G
4    AAT  r1 0.000398571  A  A  T
5    ACA  r1 0.000398571  A  C  A
6    ACC  r1 0.000398571  A  C  C
```

Model the log of the probability of Zif268 zinc finger binding to any sequence of nucleotides as a linear function of the sequence composition. Explicitly state what the covariates are. Scientists believe that, due to physical constraints of the DNA-protein interactions, positions within a binding site often interact. Evaluate this hypothesis using the available data and suggest a final model. Interpret your model in detail.

Does the constant variance assumption on the logarithm scale seem justified? One possible graphical check is

```
> library(ggplot2)
> qplot(binding, trimer, data=lKa, log="x")
> qplot(trimer, binding, data=lKa, log="y", geom="boxplot") + coord_flip()
```

5. (25 pts) A tax consultant studied the current relation between selling price and assessed valuation of one-family residential dwellings in a large tax district by obtaining data for a random sample of 16 percent "arm's length" sales transactions of one family dwellings located on corner lots and for a random sample of 48 recent sales of one-family dwellings not located on corner lots. In the collected data (available in `assessedval.txt`), both selling price Y and assessed valuation X_1 are expressed in thousand dollars, whereas lot location X_2 is coded 1 for corner lots and 0 for non-corner lots.

```
> assessed <- read.table(paste(DataURL, "assessedval.txt", sep=''),
+                        header=FALSE)
> names(assessed) <- c("Selling", "Assessed", "Corner")
> str(assessed <-
+     within(assessed, Corner <- factor(Corner, labels=c("N","Y"))))
'data.frame':        64 obs. of  3 variables:
 $ Selling : num  78.8 73.8 64.6 76.2 87.2 70.6 86 83.1 94.5 71.2 ...
 $ Assessed: num  76.4 74.3 69.6 73.6 76.8 72.7 79.2 75.6 78.1 76.9 ...
 $ Corner  : Factor w/ 2 levels "N","Y": 1 1 1 1 1 2 1 1 1 2 ...
```

Assume that the error variances in the two populations are equal and the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i$$

is appropriate.

(a) Plot the sample data for the two populations in an appropriate way. Does the regression relation appear to be the same for the two populations?

(b) Test for identity of the regression functions for dwellings on corner lots and dwellings in other locations at the significance level of 5%. State the alternatives, decision rule, and conclusion.

(c) Plot the estimated regression functions for the two populations and describe the nature of the differences between them.

(d) Find the 95% confidence interval for the estimate of β_1.

6. (25 pts) Consider the following two models where $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 I$:

   Model A: $Y = X_1\beta_1 + \epsilon$,
   Model B: $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$.

   Show that $R_A^2 \leq R_B^2$, where R^2 is defined as the multiple R-squared in the linear regression model. What does this imply for the usage of multiple R-squared in selecting among models of different dimensions?

7. (50) Suppose we want to fit the no intercept model $Y_i = \beta X_i + \epsilon_i$, $i = 1, \cdots, n$ using weighted least squares. Assume that the observations are uncorrelated but have unequal variances.

   (a) Find a general formula for the weighted least squares estimator of β.

   (b) What is the variance of the weighted least squares estimator?

   (c) Suppose that $\text{Var}(Y_i) = cX_i$, that is the variance of Y_i is proportional to the corresponding X_i. Using the results of part (a) and (b), find the weighted least squares estimator of β and the variance of this estimator.