Stat 601 Homework 4 Due 10/30/2015

1. Consider the following two models where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$:

   Model A: $Y = X_1\beta_1 + \epsilon$,

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

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

2. Suppose we want to fit the no intercept model $Y_i = X_i\beta + \epsilon_i,\ i = 1, \dots, n$ 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 $Var(Y_i) = cX_i$, 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.

3. Consider the following regression model

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

   and the transformation

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \epsilon_i$$

   where

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \qquad X_{ij}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ij} - \bar{X}_j}{s_{X_j}}\right),\ j = 1, 2,$$

   where s_Y and s_{X_j} represent the respective standard deviations.

   (a) Show that

$$Var(\hat{\beta}^* = \frac{\sigma^{*2}}{1 - r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

   where σ^{*2} is the error term variance for the transformed model and r_{12} is correlation between the predictors X_1 and X_2.

   (b) What can you say about the effect of intercorrelations among the predictor variables on the estimated regression coefficients?

4. The 6 observations made are the non-blank entries in the following incomplete two-way layout:

	B1	B2	B3
A1	y_{11}	y_{12}	
A2	y_{21}		y_{23}
A3		y_{32}	y_{33}

Consider the following Gaussian linear mode

$$y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}$$

where ϵ_{ij}s are independent normal random variable with mean 0 and standard deviation $\sigma > 0$.

(a) Let $\beta = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)'$, and $Y = (y_{11}, y_{12}, y_{21}, y_{23}, y_{32}, y_{33})'$. Find the matrix X such that

$$Y = X\beta + \epsilon$$

and rank(X).

(b) Show that components of β are not estimable.

(c) Show that $\psi_1 = \alpha_1 - \alpha_2$ and $\psi_2 = \alpha_1 + \alpha_2 - 2\alpha_3$ are both linearly estimable.

5. 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 $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, where $\mathbf{V}$ 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.

   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 $\mathbf{V}$, 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, $\mathbf{V}$, but when $\mathbf{V}$ 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 $(\beta_0, \beta_1, \beta_2, \beta_3)$ 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.

6. 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 head-width 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

```
> ants <- read.table("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.

(a) 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?

(b) How do you interpret the coefficients in this model? Be explicit.

(c) 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.

(d) 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 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?

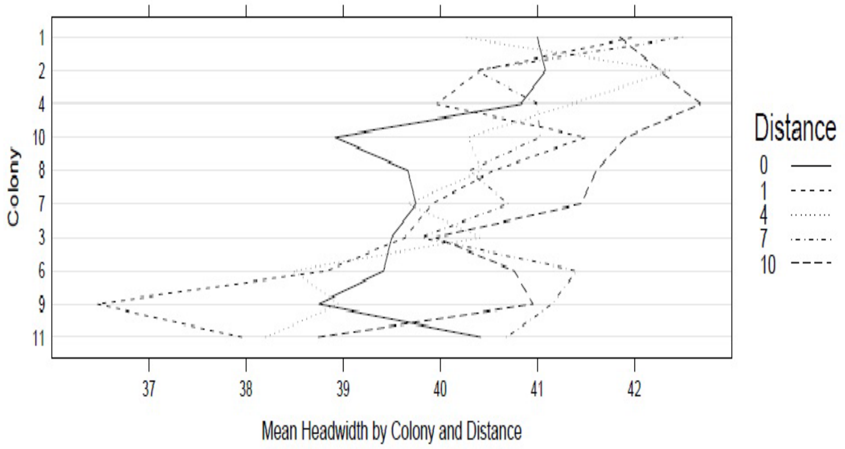

Figure 1: *Interaction plot for part (e).*

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

7. A recent research topic in the treatment of HIV is to determine whether the genotype of a patient's HIV virus can be used to decide on what type of treatment a patient should receive if the patient is failing his/her current therapy. For this purpose, a scoring system called ”Genotypic Sensitivity Score (GSS)” has been developed. The data are available in hmw3q1_data.txt on the course website. The first column represents the GSS and the second column is the patient viral load (VL), that measures the amount of virus in the blood, at a future time point.

```
> str(ql <- read.table("hmw3q1_data.txt", sep="", header=TRUE))
'data.frame': 48 obs. of 2 variables:
$ GSS: num 4 13.4 7.4 2.7 13.5 10.3 11.7 4.5 14.3 5.8 ...
$ VL : int 40406 2603 55246 22257 400 95505 5537 3205 90 12394 ...
```

(a) Plot the data. Do you think it is worthwhile testing for the presence of outliers? If yes, proceed with the test.

(b) If you identified any outliers in the above step, remove them from the data. Fit a simple linear regression model with VL as the outcome and GSS as the predictor.

(c) Does the data (possibly with outliers removed) satisfy the usual regression assumptions? Provide supporting diagnostic plots.

(d) Can you think of a transformation to apply for the linear model assumptions to be satisfied? If yes, reanalyze the data after transformation.

8. In a small scale experimental study of the relation between degree of brand liking and moisture content and sweetness of the product, data were collected on 16 subjects. These data are available in brand_preference.txt.

```
> str(br <- read.table("brand_preference.txt", sep="", header=TRUE))
'data.frame': 16 obs. of 3 variables:
$ Brand_Liking : num 64 73 61 76 72 80 71 83 83 89 ...
$ Moisture_Content: num 4 4 4 4 6 6 6 6 8 8 ...
$ Sweetness : num 2 4 2 4 2 4 2 4 2 4 ...
```

(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 moisture content interpreted?

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

(d) Prepare an added variable plot for each of the predictor variables (you might find av.plot of the cars package useful).

(e) Do your plots in part (d) suggest that the regression relationship in the fitted regression function $Y \sim\ X_1 + X_2$ (part b) are inappropriate for any of the predictor variables? Explain.

(f) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the outlier test with $\alpha = 0.10$. State the decision rule and conclusion.

(g) Are any of the observations outlying with regard to their X values?

(h) Calculate Cook's distance for each case and prepare an index plot. Are any cases influential according to this measure?

9. *[Two-way ANOVA]* A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (factors A and B) in the compound were varied at three levels each and volunteers were assigned to each of the nine treatments randomly. The data are available in hayfever.txt on the course website.

```
> str(hayfever <- within(read.table("hayfever.txt", sep=""),
+ header = TRUE),
+ {
+ A <- factor(A)
+ B <- factor(B)
+ id <- factor(id)
+ }))
```

```
'data.frame': 36 obs. of 4 variables:
$ hours: num 2.4 2.7 2.3 2.5 4.6 4.2 4.9 4.7 4.8 4.5 ...
$ A : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
$ B : Factor w/ 3 levels "1","2","3": 1 1 1 1 2 2 2 2 3 3 ...
$ id : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
```

(a) Fit the two way ANOVA model, including interactions. What is the estimated mean when factor A is 2 and factor B is 3?

(b) Using appropriate diagnostic plots, check whether there is any violation of normality.

(c) Create a plot with factor A on the x-axis, and, using 3 plotting symbols, the mean for each level of factor B above each level of factor A. Do you think there are any interactions?

(d) Test for an interaction at level $\alpha = 0.05$.

(e) Test for main efects of factors A and B.