FALL 2010
Stat 849: Homework Assignment 3
Due: October 29, 2010
Total points = 300

1. 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.

```
> DataURL <- "http://www.stat.wisc.edu/~st849-1/data/"
> str(q1 <- read.table(paste(DataURL, "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.

2. 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(paste(DataURL, "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? Comment 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?

3. *[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(paste(DataURL, "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 effects of factors A and B.

4. Verify the following properties of the residual vector $\hat{\epsilon}$.

(a) $E(\hat{\epsilon}) = 0$.

(b) $\text{cov}(\hat{\epsilon}) = \sigma^2(I - H)$.

(c) $\text{cov}(\hat{\epsilon}, Y) = \sigma^2(I - H)$.

(d) $\text{cov}(\hat{\epsilon}, \hat{Y}) = 0$.

(e) $\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0$.

(f) $\hat{\epsilon}'Y = Y'(I - H)Y$.

(g) $\hat{\epsilon}'\hat{Y} = 0$.

(h) $\hat{\epsilon}'X = 0'$.

5. 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

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \quad X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_{X_k}}\right), \quad k = 1, 2,$$

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

(a) Show that

$$\sigma^2_{\beta_1^*} = \sigma^2_{\beta_2^*} = \frac{\sigma*^2}{1 - r_{12}^2},$$

where $\sigma*^2$ 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?

6. 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 model which assumes that the observed $\{y_{ij}\}$ are independent random variables such that $y_{ij} \sim N(\eta_{ij}, \sigma^2)$ with

$$\eta_{ij} = a_i + b_j,$$

for every observed pair (i, j). The values of the real-valued parameters $\{a_i : 1 \le i \le 3\}$ and $\{b_j : 1 \le j \le 3\}$ and the value of $\sigma^2 > 0$ are unknown.

(a) Let $\beta = (a_1, a_2, a_3, b_1, b_2, b_3)'$. Let y be the vector of the observed y_{ij}, taken in order: $y = (y_{11}, y_{12}, y_{21}, y_{23}, y_{32}, y_{33})'$ and let $\eta = E(y)$ under the above linear model. Find the matrix X such that $\eta = X\beta$. Find rank(X).

(b) Consider the following Theorem:

Theorem 1 *In the model $y = X\beta + \epsilon$, where $E[y] = X\beta$ and X is $n \times p$ of rank $k < p \le n$, the linear function $\lambda'\beta$ is estimable if and only if any of the following conditions hold:*

i. λ' is a linear combination of the rows of X, that is, there exists a vector a such that

$$a'X = \lambda'.$$

ii. λ' is a linear combination of the rows of $X'X$ or λ is a linear combination of the columns of $X'X$, that is, there exists a vector r such that

$$r'X'X = \lambda' \quad or \quad X'Xr = \lambda.$$

iii. λ or λ' is such that

$$X'X(X'X)^-\lambda = \lambda \quad or \quad \lambda'(X'X)^-X'X = \lambda',$$

where $(X'X)^-$ is any generalized inverse of $X'X$.

Show that components of β are not estimable.

(c) Show that $\psi_1 = a_1 - a_2$ and $\psi_2 = a_1 + a_2 - 2a_3$ are both linearly estimable.