STAT 511 Assignment 6 Name _____
Spring 2002

Reading Assignment: Rencher: Chapter 8. Chapter 8 covers tests of hypotheses and confidence intervals for parameters in regression models. Reading Chapter 8 will help you apply the general linear model theory we are developing in Stat 511 to the analysis of multiple regression models. More information on regression analysis in given in Chapters 6, 7 and 9. This material was covered in Stat 500, and we will not repeat it in Stat 511. You can review as much of Chapters 6, 7, and 9 as you find useful. Next we will go into Chapter 14 to analyze unbalanced factorial experiments. We will consider the material on balanced factorial experiments in Chapter 13 as a special case. The material in Chapters 12 and 13 was covered in Stat 500.

Written Assignment: On-campus students: Due Wednesday, March 6, in class. Solutions will be distributed at that time. No late assignments will be accepted. Distance students: Put it in the mail or e-mail or FAX by March 14. Solutions will be posted on the course web page by 6 pm on March 14.

First Exam: The first exam will be given on Thursday, March 7, from 7-9 pm in 2245 Coover Hall. Please bring pencils, erasers, and a simple calculator. Paper and formula sheets will be provided. Distance students will be contacted their arrangements for this exam. You will need a two hour time period. A formula sheet will be posted on the course web page in the near future. Feel free to make suggestions for additions, deletions, clarifications or corrections. Previous exams and solutions have been posted on the course web page.

1. Suppose you are designing a new study of the yield of a chemical process like the one partially analyzed in problems 4 through 6 on assignment 5. Suppose the engineers assigned to your project wish to run the process at the same five temperature values for each of two new catalysts. Call them catalyst A and catalyst B. The proposed model for the observed yield when the process is run with the i-th catalyst at the j-th temperature level is

$$Y_{ijk} = \mu + \alpha_i + \beta(T_{ij} - 100) + \varepsilon_{ijk} \qquad i = 1, 2, \quad j = 1, 2, \ldots, 5, \quad \text{and} \quad k=1,\ldots,n$$

where T_{ij} is the temperature at which the process was run, and $\varepsilon_{ijk} \sim NID(0,\sigma^2)$. When the runs are made we will have n replicates for each the ten temperature/catalyst combinations. The engineers want to test the null hypothesis $H_0 : \alpha_1 = \alpha_2$ against the alternative $H_0 : \alpha_1 \neq \alpha_2$ using a type I error level of $\alpha = 0.05$. Relative to the value of the error variance, σ^2, they wish to make the number of replicates (n) large enough to have probability of at least 0.90 of rejecting the null hypothesis if $\alpha_1 - \alpha_2 = 0.5\sigma$. What is the smallest value of n that satisfies these conditions?

2. Marcuse(1949, *Biometrics,* **5**) recorded moisture content for three types of cheese made by two different methods. Two pieces of cheese were measure for each type and each method. The data are shown below.

Treatment	Moisture Content Measurements	
Type A made with Method 1	$Y_{11} = 39.02$	$Y_{12} = 38.79$
Type B made with Method 1	$Y_{21} = 35.74$	$Y_{22} = 35.41$
Type C made with Method 1	$Y_{31} = 37.02$	$Y_{32} = 36.00$
Type A made with Method 2	$Y_{41} = 38.96$	$Y_{42} = 39.01$
Type B made with Method 2	$Y_{51} = 35.58$	$Y_{52} = 35.52$
Type C made with Method 2	$Y_{61} = 35.70$	$Y_{62} = 36.04$

Consider the model $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim \text{NID}(0, \sigma^2)$, i=1,2,3,4,5,6, and j=1,2. This model can be expressed in matrix form as

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{41} \\ Y_{42} \\ Y_{51} \\ Y_{52} \\ Y_{61} \\ Y_{62} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{41} \\ \varepsilon_{42} \\ \varepsilon_{51} \\ \varepsilon_{52} \\ \varepsilon_{61} \\ \varepsilon_{62} \end{bmatrix}$$

(a) What is the distribution of $\underset{\sim}{Y} = (Y_{11}\ Y_{12}\ Y_{21}\ Y_{22}\ Y_{31}\ Y_{32}\ Y_{41}\ Y_{42}\ Y_{51}\ Y_{52}\ Y_{61}\ Y_{62})^T$?

(b) Determine which of the following are testable hypotheses. You only need to state if the hypothesis is testable or not testable.

i. $H_0 : \alpha_1 = \alpha_2 = \alpha_3$

ii. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$

iii. $H_0 : \mu + \alpha_1 = 39$ and $\mu + \alpha_4 = 39$

iv. $H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix} \underset{\sim}{\beta} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

v. $H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \underset{\sim}{\beta} = \begin{bmatrix} 0 \\ 0 \\ 38 \end{bmatrix}$

vi. $H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & -1 & 1 \end{bmatrix} \underset{\sim}{\beta} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

(c) Express the each of the following hypotheses in the form $H_0 : C\underset{\sim}{\beta} = \underset{\sim}{0}$. If the hypothesis is testable, compute the value of the corresponding F-statistic, report the degrees of freedom and the p-value for the test, and state your conclusion.

i. After averaging across the two methods of making cheese, the average moisture content is the same for all three types of cheese.

ii. For each type of cheese, the average moisture content is not affected by the method for making cheese. (This hypothesis allows the average moisture content to vary across types of cheese).

3. The following are part of the data reported by Ryan, et.al. (1976, *Jour. of Atmo. Sci.* 33) on the formation of ice crystals. The ice crystals were formed in a growth chamber maintained at a fixed temperature (-5^{o}C) and a t a fixed level of saturation of air with water. Ice crystals were harvested at various times (seconds) and the axial length (micrometers) of each ice crystal was measured. The objective was to model how mean length of ice crystals increases with time. The data file is posted as "crystals.txt" on the course web page. It contains measurements on 16 ice crystals. The first row of the file contains variable names "time" and "length". The data are also shown below.

time	length
60	18
60	21
80	25
80	28
100	30
100	29
100	33
120	36
120	34
120	28
140	32
140	35
160	38
160	30
160	37
180	37

Note that more than one ice crystal was measured at some time points. A file containing S-Plus code for assisting you in answering some the following questions is posted in the file "crystals.ssc" on the course web page. A corresponding file with SAS code is posted as "crystals.sas". SAS users should read the data from the file posted as "crystals.dat". You could also make use of the pull down menus in S-Plus to analyze these data and make graphs.

(a) Compute least squares estimates for the parameters in the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim \text{NID}(0, \sigma^2)$. This notation means that the random errors (and the observations) have normal distributions and satisfy the Gauss-Markov property. Report the estimates and their standard errors.

(b) Define

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \text{and} \quad \underset{\sim}{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{and} \quad P_X = X(X^T X)^{-1} X^T \quad \text{and} \quad P_1 = \underset{\sim}{1}(\underset{\sim}{1}^T \underset{\sim}{1})^{-1} \underset{\sim}{1}^T .$$

Here X is the model matrix for the model in part (a). What is the distribution of the quadratic form $R(\beta_1 | \beta_0) = \frac{1}{\sigma^2} \underset{\sim}{Y}^T (P_X - P_1) \underset{\sim}{Y}$?

(c) For the model in part (a), $SS_{\text{residuals}} = \frac{1}{\sigma^2} \underset{\sim}{Y}^T (I - P_X) \underset{\sim}{Y}$ has a central chi-square distribution with (n-2)=14 degrees of freedom. Define $MS_{\text{residuals}} = \frac{SS_{\text{residuals}}}{n-2}$ and use the results from part (b) to derive the distribution of $F = \frac{R(\beta_1 | \beta_0)}{MS_{\text{residuals}}}$. Report degrees of freedom and a formula for the noncentrality parameter.

(d) What is the null hypothesis associated with the F statistic in part (c)? Justify your answer by showing that the noncentrality parameter in part (c) is zero if and only if the null hypothesis is true.

(e) Report the value of the test statistics in part (c) and state your conclusion.

(f) Examine the plot of the estimated line and observations and the residual plots provided by the code posted on the course web page. What do these plots suggest?

4. Suppose the model proposed in part (a) of problem 2 is incorrect. In particular, suppose that the correct model is $Y_i = \lambda_0 + \lambda_1 X_i + \lambda_2 X_i^2 + \eta_i$ where $\eta_i \sim \text{NID}(0, \omega^2)$. This model can be expressed in matrix notation as $\underset{\sim}{Y} = Z\underset{\sim}{\lambda} + \underset{\sim}{\eta}$, where

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad Z = \begin{bmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{bmatrix} = \left[X \,\middle|\, \underset{\sim}{d} \right] \qquad \underset{\sim}{d} = \begin{bmatrix} X_1^2 \\ X_2^2 \\ \vdots \\ X_n^2 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix}.$$

Assuming that this is the correct model for ice crystal growth, find

(a) The distribution of $\underset{\sim}{Y}^T (I - P_X) \underset{\sim}{Y}$, where X is the model matrix from problem 2, and

(b) The distribution of $\underset{\sim}{Y}^T (P_X - P_1) \underset{\sim}{Y}$.

(c) Does $\dfrac{Y^T (P_X - P_1) Y}{Y^T (I - P_X) Y / (n-2)}$ have an F-distribution? Explain.

5. Now, suppose that the model in problem 2 is correct, i.e. $\lambda_3 = 0$ and $\eta_i \sim \text{NID}(0, \sigma^2)$ for the model in problem 3.

(a) Find the distribution of $\underset{\sim}{Y}^T (I - P_Z) \underset{\sim}{Y}$, where $P_Z = Z(Z^T Z)^{-1} Z^T$ and Z is the model matrix from problem 3.

(b) Does $\dfrac{Y^T (P_X - P_1) Y}{Y^T (I - P_Z) Y / (n-3)}$ have an F-distribution when the model in problem 2 is correct? Explain.

6. Problems 3 and 4 illustrate some of the consequences of incorrectly specifying the model. When you have replication at some sets of values of the explanatory variables, as we do for the ice crystal data, you can construct a lack-of-fit test for a proposed model. We will apply a lack-of-fit test to the quadratic model from problem 3. Consider the larger model

$$Y_{ij} = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \alpha_j + \varepsilon_{ij}, \qquad \text{where } \varepsilon_{ij} \sim \text{NID}(0, \tau^2)$$

and Y_{ij} denotes the observed axial length for the j-th ice crystal measure at time X_i. This model can be expressed in matrix notation as $\underset{\sim}{Y} = W\underset{\sim}{\gamma} + \underset{\sim}{\varepsilon}$, where

$$\underset{\sim}{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \\ Y_{33} \\ Y_{41} \\ Y_{42} \\ Y_{43} \\ Y_{51} \\ Y_{52} \\ Y_{61} \\ Y_{62} \\ Y_{63} \\ Y_{71} \end{bmatrix} \qquad W = \begin{bmatrix} 1 & X_1 & X_1^2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & X_1 & X_1^2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & X_2 & X_2^2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & X_2 & X_2^2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & X_3 & X_3^2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & X_3 & X_3^2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & X_3 & X_3^2 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & X_4 & X_4^2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & X_4 & X_4^2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & X_4 & X_4^2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & X_5 & X_5^2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & X_5 & X_5^2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & X_6 & X_6^2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & X_6 & X_6^2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & X_6 & X_6^2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & X_7 & X_7^2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad \underset{\sim}{\gamma} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \end{bmatrix}$$

Note that the first three columns of W comprise the Z matrix from problem 3. Define $P_Z = Z(Z^TZ)^{-1}Z^T$ and $P_W = W(W^TW)^{-}W^T$.

(a) Consider the test statistic

$$F = \frac{Y^T(P_W - P_Z)Y/(7-3)}{Y^T(I - P_W)Y/(n-7)}.$$

Report a formula for the noncentrality parameter for the distribution of this statistic and use it to show that this is an appropriate lack-of-fit test.

(b) Would it be better to use $Y^T(I - P_W)Y/(n-7)$ in the denominator of this test statistic instead of $Y^T(I - P_Z)Y/(n-3)$? Explain.