STAT 511 Assignment 5 Name _____
Spring 2002

Reading Assignment: Rencher: Read Chapters 11, 4 and 5. This is the same set of readings given on assignment 4. Chapter 11 covers the basics on estimable functions and testable hypotheses. Chapter 4 reviews some properties of the multivariate normal distribution, and Chapter 5 considers distributions of quadratic forms and related F and t distributions. The introduction to regression analysis in Chapters 6 and 7 and the regression diagnostics in Chapter 9 were covered in Stat 500, and we will not repeat the coverage of this material in Stat 511. You can review as much of this material as you find useful. 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.

Written Assignment: On-campus students: Due Wednesday, February 28, in class.
Distance students: Put it in the mail or e-mail or FAX by March 1.

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 should plan to take this exam around March 15. You will need a two hour time period. The 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 $\underset{\sim}{Y} \sim N\left(\underset{\sim}{m}, \Sigma\right)$ where

$$\underset{\sim}{\mu} = \begin{bmatrix} -1 \\ 0 \\ -3 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 0.75 & 0.00 & 0.25 \\ 0.00 & 1.00 & 0.00 \\ 0.25 & 0.00 & 0.75 \end{bmatrix}$$

Let

$$A = \begin{bmatrix} 1.5 & 0.0 & -0.5 \\ 0.0 & 1.0 & 0.0 \\ -0.5 & 0.0 & 1.5 \end{bmatrix}$$

Make use of S-PLUS and formulas given in Results 4.6a and 4.6b in the course notes for the mean and variance of a quadratic form to answer the following questions:

(a) Evaluate $E\left(\underset{\sim}{Y}^T A \underset{\sim}{Y} \right)$

(b) Evaluate $Var\left(\underset{\sim}{Y}^T A \underset{\sim}{Y} \right)$

(c) Does $\underset{\sim}{Y}^T A \underset{\sim}{Y}$ have a chi-square distribution? Explain. (Use S-Plus to check the condition of Result 4.7.)

(d) If $\Sigma = Var(\underset{\sim}{Y})$ was $\sigma^2 I$ instead of the matrix shown above, would $\underset{\sim}{Y}^T A \underset{\sim}{Y} / s^2$ have a chi-square distribution? Explain.

2. Suppose $Y_1, Y_2, ..., Y_n$ are a simple random sample (this means that the observations are independent of each other and they all have the same distribution) from a population of values with a normal distribution with population mean μ and population variance σ^2. Define $\underset{\sim}{Y} = (Y_1, ..., Y_n)^T$. Then,

$$\underset{\sim}{Y} \sim N\left(\mu \underset{\sim}{1}, \sigma^2 I\right)$$

where $\underset{\sim}{1}$ is a vector of ones. Let $P_1 = \underset{\sim}{1}(\underset{\sim}{1}^T \underset{\sim}{1})^{-1} \underset{\sim}{1}^T$.

(a) The sample variance, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2$, can be expressed as

$$S^2 = \frac{1}{n-1} \underset{\sim}{Y}^T (I - P_1) \underset{\sim}{Y}$$

Use results 4.6a and 4.6b in the course notes to find $E(S^2)$ and $Var(S^2)$.

(b) Use result 4.7 in the course notes to show that $(n-1)S^2/\sigma^2$ has a central chi-squared distribution for any value of μ. What are the degrees of freedom? (Note that this result also provides the mean and variance of S^2.)

(c) Use result 4.8 in the course notes to show that $n\overline{Y}^2 = \underset{\sim}{Y}^T P_1 \underset{\sim}{Y}$ is distributed independently of $S^2 = \frac{1}{n-1} \underset{\sim}{Y}^T (I - P_1) \underset{\sim}{Y}$.

(d) Find the distribution of

$$V = \frac{n\,\overline{Y}^{\,2}}{S^2}\ .$$

(e) By examination of the non-centrality parameter for its distribution, determine the hypothesis that can be tested with the V statistic in part (d).

3. Consider the linear model $\mathbf{Y} = \mathbf{X}\,\mathbf{b} + \mathbf{e}$ with

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{11} \\ \mathbf{Y}_{12} \\ \mathbf{Y}_{21} \\ \mathbf{Y}_{22} \\ \mathbf{Y}_{23} \\ \mathbf{Y}_{24} \\ \mathbf{Y}_{31} \\ \mathbf{Y}_{32} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} m \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

and $\mathbf{e} \sim N(0, \sigma^2 I)$.

(a) Determine which of the following hypotheses are testable. Just report a yes or no answer.

(i) H_0: $\alpha_1 = \alpha_2$
(ii) H_0: $\alpha_1 - 2\alpha_2 + 3\alpha_3 = 0$
(iii) H_0: $\alpha_3 = 0$
(iv) H_0: $\mu = 0$
(v) H_0: $\alpha_1 = \alpha_3$ and $\alpha_1 - 2\alpha_2 + \alpha_3 = 0$
(vi) H_0: $\alpha_1 = \alpha_2$ and $\alpha_1 = \alpha_3$ and $2\alpha_1 - \alpha_2 - \alpha_3 = 0$

(b) Show how to construct a test of the null hypothesis H_0: $C\hat{\mathbf{a}} \equiv \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

against the alternative H_A: $\alpha_1 \neq \alpha_2$ or $\alpha_1 - 2\alpha_2 + \alpha_3 \neq 0$. You can express your answer using formulas for quadratic forms. You answer should consist of the following parts:

(i) Show that the quadratic form in the denominator of your F-statistic,

$$\frac{1}{\sigma^2}\text{SSE} = \frac{1}{s^2}\,\mathbf{Y^T(I - P_X)Y}\,, \quad \text{where} \quad P_X = X(X^TX)^{-}X^T,$$

has a central chi-square distribution with 5 degrees of freedom.

(ii) Express the numerator of your F-statistic as a quadratic form and show that it is distributed independently of the SSE in part (i). (Use result 4.8 from the notes.)

(iii) Show that the quadratic form identified in part (ii) is distributed as a multiple of a non-central chi-square distribution. (Use result 4.7 from the course notes.)

(iv) Use the results from parts (i), (ii) and (iii) and appeal to the definition of the non-central F-distribution to show that your F-statistic has a non-central F-distribution. Report degrees of freedom and express the non-centrality parameter as a function of α_1, α_2, α_3.

(v) Show that your test statistic has a central F-distribution when the null hypothesis is true.

(c) Consider an F-test of the null hypothesis H_0: $\alpha_1=\alpha_2=\alpha_3=0$ against the alternative H_A: $\alpha_1 \neq \alpha_2$ or $\alpha_1 \neq \alpha_3$ or $\alpha_2 \neq \alpha_3$. Explain why testing this null hypothesis is equivalent to testing the null hypothesis in part (b).

4. Consider the study of the yield of a chemical process reported in Problem 2 on Assignment 2. One model for the observed yield when the process is run with the i-th catalyst at the j-th temperature level is

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

where

Y_{ij} = the observed yield for the run using the i-th catalyst and the j-th temperature level.

α_i is associated with the i-th catalyst

T_{ij} = the temperature under which the process was run.

and $\varepsilon_{ij} \sim \text{NID}(0,\sigma^2)$. Here we have included the assumption of a normal distribution for the random errors. This model is written in matrix form as

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \\ Y_{25} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & -10 \\ 1 & 1 & 0 & -5 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 5 \\ 1 & 1 & 0 & 10 \\ 1 & 0 & 1 & -10 \\ 1 & 0 & 1 & -5 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 5 \\ 1 & 0 & 1 & 10 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}$$

(a) Write the null hypothesis H_0: $E(Y_{ij})$ = constant, for all (i,j), in the form H_0: $C\mathbf{b} = \mathbf{0}$ where $\mathbf{b} = (\mu\ \alpha_1\ \alpha_2\ \beta)^T$ is the parameter vector.

(b) Show that the null hypothesis in part (a) is testable.

(c) Present a formula for $\mathbf{SS_{H_0}}$, corresponding to the null hypothesis in part (a), and show that it is distributed as a multiple of a central chi-square random variable when the null hypothesis is true.

(d) Let $\mathbf{e} = (I\text{-}P_X)\mathbf{Y}$ denote the vector of residuals for this model. Show that

$$\frac{1}{\sigma^2}\mathbf{SSE} = \frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} = \frac{1}{\sigma^2}\mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y}$$

has a central chi-square distribution and report its degrees of freedom.

(e) Show that $\mathbf{SS_{H_0}}$ and SSE are independent.

(f) Show how to construct an F-test for the null hypothesis in part (a). Report the degrees of freedom and give a formula for the non-centrality parameter.

(g) Using the data from problem 2 on assignment 4, evaluate your F-statistic from part (f), compute a p-value and state your conclusion.

5. Consider a normal theory Gauss-Markov model for the chemical process data from Problem 2 on Assignment 4. Let $\mathbf{Y} \sim N(W\mathbf{g}, \sigma^2 I)$ where

$$\mathbf{W}g = \begin{bmatrix} 1 & 1 & -10 \\ 1 & 1 & -5 \\ 1 & 1 & 0 \\ 1 & 1 & 5 \\ 1 & 1 & 10 \\ 1 & -1 & -10 \\ 1 & -1 & -5 \\ 1 & -1 & 0 \\ 1 & -1 & 5 \\ 1 & -1 & 10 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}$$

(a) Show that this model (call it Model II) is a reparameterization of the model (call it model I) in Problem 4 on this assignment.

(b) With respect to the yield of the chemical process and the catalyst and temperature factors, how would you interpret the parameters, γ_1, γ_2 and γ_3 in Model II?

(c) The observed yields are given on the data file for problem 2 on assignment 4. Use S-PLUS to compute the following for Model II.

(i) The OLS estimator for $\mathbf{g}$, call it $\hat{\mathbf{g}}$,

(ii) The OLS estimator $\hat{\mathbf{Y}} = \mathbf{W}\hat{\mathbf{g}}$ for $E(\mathbf{Y}) = W\tilde{\mathbf{a}}$.

(iii) The vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$.

(iv) $SSE = e^T e = \mathbf{Y}^T(I - P_W)Y$, where $P_W = W(W^T W)^{-1} W^T$.

(v) $\mathbf{R}(g_1) = \mathbf{Y}^T \mathbf{P}_{\mathbf{W}_1} \mathbf{Y}$ where $W_1 = 1$ is the first column of W and $P_{W_1} = W_1(W_1^T W_1)^{-1} W_1^T$.

(vi) $\mathbf{R}(g_2 \mid g_1) = \mathbf{Y}^T(\mathbf{P}_{\mathbf{W}_2} - \mathbf{P}_{\mathbf{W}_1})\mathbf{Y}$ where W_2 contains the first two columns of W and $P_{W_2} = W_2(W_2^T W_2)^{-1} W_2^T$.

(vii) $R(\mathbf{g}_3 \mid \mathbf{g}_1\ \mathbf{g}_2) = \mathbf{Y}^{\mathbf{T}}(\mathbf{P}_{\mathbf{W}} - \mathbf{P}_{\mathbf{W}_2})\ \mathbf{Y}$.

(viii) Collect the sums of squares from parts (iv), (v), (vi), and (vii) into an ANOVA table. Include degrees of freedom, means squares and values of F-tests.

(d) Use Cochran's theorem to make a statement about the distributions of the sums of squares in Parts (vi), (v), (vi) and (vii) of Part (c).

(e) Report a formula for the non-centrality parameter of the non-central F distribution of

$$F = \frac{R(g_3 \mid g_1\ g_2)}{SSE/(n-3)}$$

Use it to argue that this statistic provides a test of the null hypothesis $H_0: \gamma_3 = 0$ against the alternative $H_A: \gamma_3 \neq 0$.

(f) Report a formula for the non-centrality parameter of the non-central F distribution of

$$F = \frac{R(\gamma_2 \mid \gamma_1)}{SSE/(n-3)}$$

Use it to identify the null and alternative hypotheses associated with this test statistic.

(g) Estimate the covariance matrix for $\hat{\mathbf{g}}$, the OLS estimator for $\mathbf{g}$, and use it to obtain 95 percent confidence intervals for

i. γ_2

ii. The mean yield when the process is run at 120 degrees C with catalyst A. (Note that the temperature value of 120C corresponds to 20=120-100 for this model.)

(h) Construct a 95% confidence interval for the error variance.

You could numerically verify that Models I and II are equivalent by using S-PLUS to compute the following for Model I. (Do not submit these results.)

(i) The OLS estimator $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = P_X\mathbf{Y}$ for $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\mathbf{b}$, where $P_X = X(X^TX)^- X^T$.

(ii) The vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (I - P_X)\mathbf{Y}$.

(iii) $SSE = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(I - P_X)\mathbf{Y}$

(iv) $R(\mu) = \mathbf{Y}^T P_{X_1} \mathbf{Y}$ where X_1 is the first column of X and $P_{X_1} = X_1(X_1^T X_1)^{-1} X_1^T$.

(v) $\mathbf{R}(a_1, a_2 \mid m) = \mathbf{Y}^T(\mathbf{P}_{X_2} - \mathbf{P}_{X_1})\mathbf{Y}$ where X_2 consists of the first three columns of X and $P_{X_2} = X_2(X_2^T X_2)^{-1} X_2^T$.

(vi) $\mathbf{R}(b \mid m, a_1, a_2) = \mathbf{Y}^T(\mathbf{P}_X - \mathbf{P}_{X_2})\mathbf{Y}$

Note that these sums of squares provide the same ANOVA table obtained in Part (d) for Model II.

6. (a) Note that Model II in Problem 5 is obtained from Model I in problem 4 through a reparameterization that places one linear restriction on the parameters in Model I. What is the restriction?

(b) Show that the linear combination of parameters involved in the restriction identified in Part (a) is not an estimable function of the parameters in Model I?

Problems 4, 5 and 6 illustrate that the use of a generalized inverse can be avoided by putting enough linear restrictions on parameters to create a model matrix of full column rank. This is true for any linear model. The restrictions placed on the parameters must correspond to non-estimable functions. Of course the choice of restrictions is not unique and different sets of restrictions will lead to different interpretations and different least squares estimates of non-estimable functions of parameters, but estimable functions of parameters will be invariant to the choice of restrictions.

7. Consider another model for the chemical process data:

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 0 & -10 & 0 \\ 1 & 1 & 0 & -5 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 5 & 0 \\ 1 & 1 & 0 & 10 & 0 \\ 1 & 0 & 1 & 0 & -10 \\ 1 & 0 & 1 & 0 & -5 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 5 \\ 1 & 0 & 1 & 0 & 10 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_2 \\ \beta_3 \\ \\ \end{bmatrix} + \varepsilon$$

where $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$. Is this model a reparameterization of the models in problems 4 and 5? Explain.

8. 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 this assignment. 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?