STAT 511 Assignment 4 Name _____
Spring 2002

Reading Assignment: Rencher: Read Chapters 11, 4 and 5. Chapter 11 was part of an earlier reading assignment. It 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 Friday, February 15, in class.

Distance students: Put it in the mail or e-mail or FAX by February 22.

1. A food scientist performed the following experiment to study the effects of combining three different fats and three different surfactants on the specific volume of bread loaves. Four batches of dough were made for each of the six combinations of fat and surfactant. Ten loaves of bread were made from each batch of dough and the average volume of the ten loaves was recorded for each batch. Unfortunately, some of the yeast used to make some batches of dough was ineffective and data from the loaves made from those batches had to be removed from the analysis. Fortunately, all six combinations of the levels of fat and surfactant were observed at least once. The data (average volume of 10 loaves) are shown below.

		Surfactant	
	A	**B**	**C**
Fat 1	6.7 4.3 5.7	7.1 5.9 5.6	5.5 6.4 5.8
Fat 2	5.9 7.4 7.1	5.6 6.8	6.4 5.1 6.2 6.3

Consider the model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ where $\varepsilon_{ijk} \sim NID(0, \sigma^2)$ and Y_{ijk} denotes the average of the volumes of ten loaves of bread made from the k-th batch of dough using the i-th fat and the j-th surfactant. For each of the following linear functions of the parameters, determine if it is estimable. If it is estimable, give a vector $\mathbf{a}$ such that $E(\mathbf{a}^T\mathbf{Y})$ is equal to that linear combination of parameters. If it is estimable, describe what that linear combination of parameters represents with respect to the effects of the fats and surfactants on mean bread volume.

(a) μ

(b) α_2

(c) $\beta_2 - \beta_3$

(d) γ_{23}

(e) $\mu + \alpha_2 + \beta_3 + \gamma_{23}$

(f) $\gamma_{11} - \gamma_{12}$

(g) $\gamma_{11} - \gamma_{13} - \gamma_{21} + \gamma_{23}$

(h) $(\beta_2 - \beta_3) + \frac{1}{2}(\gamma_{12} + \gamma_{22} - \gamma_{13} - \gamma_{23})$

(i) $\gamma_{12} + \gamma_{22} - \gamma_{13} - \gamma_{23}$

2. Data were collected to study the effect of temperature on the yield of a chemical process. Two different catalysts A and B, were used in the study. Yields were measured under 5 different temperatures for each catalyst. The data are as follows:

Run	**Yield (grams) Y**	**Temperature (°C) T**	**Catalyst**
3	20	90	A
10	24	95	A
4	27	100	A
8	33	105	A
5	38	110	A
9	25	90	B
2	29	95	B
6	32	100	B
1	37	105	B
7	41	110	B

Each run can be considered as an independent observation. The order in which the runs were made was randomized.

Consider the linear model

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

where

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

α_i corresponds to the $i-$th catalyst

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

(a) For this linear model the vector of mean responses can be written as $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\,\mathbf{b}$. Write out the model matrix $\mathbf{X}$ corresponding to the parameter vector $\hat{\mathbf{a}} = (\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \beta)^T$.

(b) Determine which, if any, of the following quantities are estimable. For each estimable quantity, report the value of a vector $\mathbf{a}$ such that $\mathbf{a}^T\mathbf{Y}$ satisfies the definition of an estimable function of $\mathbf{b}$.

(i) μ

(ii) $\mu + \alpha_2$

(iii) β

(iv) $\alpha_1 - \alpha_2$

(v) $\mu + \beta\, T$ where T is any specified temperature

(vi) $\mu + \alpha_1 + \beta\,(T-100)$ where T is any specified temperature

(c) The data are posted in the file hw402p2.txt on the course web page. This file has five columns. The first two columns match the first two columns in the table shown above. The third and fourth column use dummy variables to indicate which catalyst was used. The third column is coded "1" when catalyst A was used and coded "0" when the other catalyst was used. The fourth column is coded "1" when catalyst B was used and coded "0" when the other catalyst is used. The fifth column contains the temperature values minus 100. Use the command

```
W  <-   read.table("c:/stat511/hw402p2.txt",header= T)
```

to enter these data into a data frame in S-PLUS. Of course, you should replace

"c:/stat511/hw402p2.txt" with the name of the file in which you stored these data.

Use the command

```
Y <- as.matrix(W[ ,2 ])
```

to create a vector of observed responses. Use the command

```
X  <- as.matrix(cbind(rep(1,length(Y)),W[ ,3:5]))
```

to construct the model matrix for the model in part (a).

If you are using S-PLUS version 6 for Windows or the UNIX version 5.1 of S-PLUS (available on the system of VINCENT workstations at ISU), use the ginverse( ) function to compute a generalized inverse of $\mathbf{X}^T\mathbf{X}$ and report the result**.** If you are using an older version of S-PLUS for windows, the ginverse( ) function will not be available. Those users can use the ginv( ) function in the MASS library of functions. The MASS library comes with any Windows version of S-PLUS and you downloaded it onto you computer when you installed S-PLUS. To use a function in the MASS library, you must first establish a path to the library by issuing the command

```
library(MASS)
```

from the command window. Then when you issue the command

```
M <- ginv(t(X)%*%X)
```

S-PLUS will looking the MASS library for the ginv( ) function. If you fail to issue the library(MASS) command, S-PLUS will only look for the ginv( ) in the default library of functions.

(d) Use S-PLUS to check if the generalized inverse computed in part (c) satisfies the four properties of the Moore-Penrose inverse. State your conclusion.

(e) Use the generalized inverse from part (c) to compute a solution to the normal equations $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$. Report your value of $\mathbf{b}$.

(f) To visually check if the proposed model is reasonable for these data and to practice

using S-PLUS to make plots, create a scatter plot of yield (Y) versus temperature (T). Use an open circle for the 5 observations with catalyst A and a filled circle for the five observations with catalyst B. Also include two parallel lines on the plot corresponding to the least squares estimates of the models for catalysts A and B. This can be done with the following code.

```
#  Splus code for problem 2 on hw4, 2002
#  Enter the data into a data frame

   W <- read.table("c:/stat511/hw402p2.txt",header=T)

# Create the response vector

   Y <- as.matrix(W[,2])

#  Construct a vector of temperatures

   temp <- X[ ,4]

# Construct a factor to represent the groups

   groups <- as.factor(W[ ,3]+2*W[ ,4])


#  Use the lm( ) function to fit the model

   lm.out <- lm(Y~groups+temp)

#  Plot the observations on the same plot
#  with the estimated lines.
#  First construct the axes for the plot
#  and print titles and labels.

   par(fin=c(7,7),mar=c(5,5,4,2),cex=1.5,mkh=1.3)
   plot(c(min(temp),max(temp)), c(min(Y),max(Y)),
   xlab='Temperature-100',ylab='Yield',
   type="n", main="Problem 2 on Assignment 4 ")

#  Now insert the observations and fitted lines

      pc <- c(1,16)
      lt <- c(1,7)
      N <- seq(1,length(groups))

      for(i in sort(unique(groups))) {
         j <- N[groups == i]
         points(temp[j], Y[j], pch = pc[i])
         lines(temp[j],lm.out$fitted[j],lty=lt[i])
      }
```

Submit your plot and comment on the results.

(g) Using your solution $\mathbf{b} = (\hat{\mu}\ \ \hat{\alpha}_1\ \ \hat{\alpha}_2\ \ \hat{\beta})^T$ to the normal equations from part (c), estimates of the lines that describe how the mean yield changes with changes in temperature are

$\hat{Y}_{1j} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta} T_{1j}$ when catalyst A is used

and

$\hat{Y}_{2j} = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta} T_{2j}$ when catalyst B is used

Would these estimates change if you used a different solution to the normal equations? Explain.

3. Consider the "common mean" model $Y_{ij} = \mu + \varepsilon_{ij}$, i = 1, 2 and j = 1, 2, ..., 5, for the data in problem 2. This model can be expressed as $\mathbf{Y} = \mathbf{1}\mu + \varepsilon$, where

$\mathbf{Y} = (Y_{11}\ \ Y_{12}\ \ Y_{13}\ \ Y_{14}\ \ Y_{15}\ \ Y_{21}\ \ Y_{22}\ \ Y_{23}\ \ Y_{24}\ \ Y_{25})^T$ is vector of observed yields and

$\mathbf{1} = (1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1)^T$ is a model matrix with one column.

(a) Note that $\hat{\mu} = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y}$ is the unique solution to the normal equations for this model. Show that $\hat{\mu} = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y} = \overline{Y}_{\bullet\bullet}$, the sample mean of the ten observations.

(b) Show that $\mathbf{P_1} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ satisfies the definition of an idempodent matrix.

(c) Give a formula for the vector of estimated means, $\hat{\mathbf{Y}} = \mathbf{P_1}\mathbf{Y}$, as a function of $\overline{Y}_{\bullet\bullet}$.

(d) Show that the uncorrected total sum of squares can be partitioned as

$SS_{\text{total, uncorrected}} = \mathbf{Y}^T\mathbf{Y} = \mathbf{Y}^T\mathbf{P_1}\mathbf{Y} + \mathbf{Y}^T(\mathbf{I} - \mathbf{P_1})\mathbf{Y} = SS_{\text{model, uncorrected}} + SS_{\text{residuals}}$.

(e) For the "common mean" model, $SS_{residuals} = \mathbf{Y^T(I - P_1)Y} = \sum_{i=1}^{2}\sum_{j=1}^{5}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$, is the quantity that one generally calls the corrected total sum of squares. Obtain a formula for $SS_{model,uncorrected} = \mathbf{Y^T P_1 Y}$ as a function of $\overline{Y}_{\bullet\bullet}$.

(f) Show that the uncorrected total sum of squares can also be partitioned as $SS_{total,\ uncorrected} = \mathbf{Y^T Y = Y^T P_1 Y + Y^T(P_X - P_1)Y + Y^T(I - P_X)Y}$, where $\mathbf{P_X}$ is the projection matrix for the model in problem 2.

(g) Show that the corrected model sum of squares for the model in problem 2 is $SS_{model,problem\ 2} = \mathbf{Y^T(P_X - P_1)Y} = SS_{residuals,common\ means\ model} - SS_{residuals,\ problem\ 2}$

the difference between the residual sum of squares for the "common means" model and the model in problem 2. The degrees of freedom for this sum of squares is the difference in the dimensions of the residual spaces for the two models, i.e., (10-1)-(10-3)=2 d.f..