STAT 511 Assignment 3 Name _____
Spring 2002

Reading Assignment: Rencher: Review Sections 2.8 and 2.13. You should have already read Chapter 11 which provides an overview of estimation, estimability, and testing hypotheses in linear models. Applications of these ideas to parameter estimation in one-way analysis of variance and regression analysis are given in Sections 12.1-12.3 and 7.1-7.5, respectively. You covered this material in Stat 500, but it may be helpful to review it at this time. The later parts of Chapters 7 and 12 are on hypothesis testing. We will need to cover the material in Chapters 4 and 5 before we begin our discussion of hypothesis testing. You should next read Chapters 4 and 5.

Written Assignment: On-campus students: Due Wednesday, February 6.
Distance students: Put it in the mail by February 14.

1. Suppose P is a non-singular matrix. Use the definitions of positive definite and positive semi-definite matrices to prove the following results.

   (a) If A is positive semidefinite, then P^TAP is positive semidefinite.

   (b) If A is positive definite, then P^TAP is positive definite.

2. Only square, nonsingular matrices have inverses, but every matrix has a generalized inverse. For example, let

$$A = \begin{bmatrix} 1 \\ 2 \\ 5 \\ -2 \end{bmatrix}$$

   (a) Show that $B = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$ is a generalized inverse for A.

   (b) Find two other generalized inverses for A.

3. Let $X = (x_1, x_2, \ldots, x_n)^T$ denote any non-zero vector of length n.
   (a) Show that X is an eigenvector of the matrix $I - X(X^TX)^{-1}X^T$.
   (b) What is the eigenvalue associated with X?
   (c) Show that any vector V that is orthogonal to X, i.e. $X^TV = 0$, is also an eigenvector of $I - X(X^TX)^{-1}X^T$.
   (d) What are the eigenvalues of $I - X(X^TX)^{-1}X^T$?

4. Let A be an nxn symmetric matrix with rank (A) = r. Here r may be smaller than n. Let

$$A = L \begin{bmatrix} \Delta_r & 0 \\ 0 & 0 \end{bmatrix} L^T$$

represent the spectral decomposition of A. Then, Δ_r is an r×r diagonal matrix containing the positive eigenvalues of A, and L is an n×n orthogonal matrix where the columns are eignenvectors of A. Show that $G = L \begin{bmatrix} \Delta_{\mathbf{r}}^{-1} & 0 \\ 0 & 0 \end{bmatrix} L^T$ satisfies the definition of the Moore-Penrose inverse of A.

5. (a) Use the eigen ( ) function in S-Plus to compute the eigenvalues and eigenvectors of

$$V = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 5 & -1 \\ 1 & -1 & 3 \end{bmatrix}$$

(b) Write an S-Plus function to compute the inverse square root matrix of a symmetric positive definite matrix. (If V is a symmetric positive definite matrix, find a matrix $V^{-1/2}$ such that $V^{-1/2}\ V^{-1/2} = V^{-1}$.) Submit a listing of your code.

(c) Use your function from part (b) to evaluate $V^{-1/2}$ for the matrix in part (a).

(d) Suppose $\underset{\sim}{Y} = (Y_1\ Y_2\ Y_3)^T$ is a random vector with $E(\underset{\sim}{Y}) = \underset{\sim}{0}$ and $VAR(\underset{\sim}{Y}) = V$, where V is the matrix in part (a). Let $\underset{\sim}{Z} = V^{-1/2}\ \underset{\sim}{Y}$. Find $E(\underset{\sim}{Z})$ and $Var(\underset{\sim}{Z})$.

6. Consider a linear model $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i$, where ε_i is a random error with $E(\varepsilon_i) = 0$ for all i = 1, 2, ..., n. This model can be expressed as a linear model with $E(Y) = X\mathbf{b}$ and $Var(\mathbf{Y}) = \Sigma$. An ordinary least squares estimator for $\mathbf{b}$ is any vector $\mathbf{b}$ that minimizes the sum of squared residuals, i.e., minimizes

$$\sum_{\mathbf{i}=1}^{\mathbf{n}} (Y_i - b_1 X_{1i} - \cdots - b_k\ X_{ki})^2 = \left(\mathbf{Y} - \mathbf{X}\,\mathbf{b}\right)^{\mathbf{T}} \left(\mathbf{Y} - \mathbf{X}\,\mathbf{b}\right).$$

As shown in the lectures, setting first partial derivatives of this quantity equal to zero yields the normal equations

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}\,\mathbf{b} = \mathbf{X}^{\mathbf{T}}\,\mathbf{Y}\ .$$

(a) Show that $\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathbf{T}}\,\mathbf{Y}$ is a solution to the normal equations for any generalized inverse $\left(X^{T}X\right)^{-}$ of $\mathbf{X}^{\mathbf{T}}\mathbf{X}$.

(b) Can every solution to the normal equations be written as $\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathbf{T}}\,\mathbf{Y}$ for some generalized inverse $\left(X^{T}X\right)^{-}$ of $X^{T}X$? Consider $\mathbf{b}^{*} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathbf{T}}\mathbf{Y} + \mathbf{a}$. Are there any non-zero vectors $\mathbf{a}$ for which $\mathbf{b}^{*} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-}\mathbf{X}^{\mathbf{T}}\mathbf{Y} + \mathbf{a}$ is a solution to the normal equations? Explain.

(c) Show that $\mathbf{b} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$ is the unique solution to the normal equations when $\mathbf{X}$ has full column rank.

7. In this problem we will use line commands to enter data into S-PLUS as a data frame, make some graphs and compute least squares estimates of parameters in a regression model. The data for this problem are stored in the file **biomass.txt** that is available from the course web page. These data were obtained from a study of soil characteristics on aerial biomass production of the marsh grass *Spartina alterniflora*, in the Cape Fear Estuary of North Carolina. (Rick A. Linthurst (1979) *Aeration, nitrogen, pH, and salinity as factors affecting Spartina alterniflora growth and dieback.* Ph.D. thesis, North Carolina State University.) Each line of this data file corresponds to a different sample. There are eight entries on each line corresponding to the following quantities in the order listed.

Location

Type of Spartina vegetation: (revegetated areas, short grass areas, Tall grass areas)

Y = aerial biomass (gm^{-2})

X_1 = soil salinity (o/oo)

X_2 = soil acidity as measured in water (pH)

X_3 = soil potassium (ppm)

X_4 = soil sodium (ppm)

X_5 = soil zinc (ppm)

The first line of this file has the variable names. (If you want to analyze these data with the SAS package you should use the data file posted under biomass.dat. There are no variable names in that file.) Copy the data file to your Stat 511 directory. Start S-PLUS on Vincent or start up a windows version of S-PLUS and get into the Command window. You can enter these data into S-PLUS as a data frame with the command

```
biomass  ←  read.table("filename",header=T)
```

Then, use the command

```
biomass
```

to view the data frame. It should have eight columns and 45 rows. Now create two matrices that will be used to fit a regression model to these data. Create a vector **Y** from the third column of the data frame and a matrix X from the last five columns of the data frame with the following commands:

```
Y ← as.matrix(biomass[ ,3])
X ← as.matrix(biomass[ ,4:8])
```

Note the use of `[ ]` to select columns from the data frame. Here, the function `as.matrix` is used to create a matrix from one or more columns of the data frame. To add a column of ones to the model matrix, use the commands

```
X0 ← rep(1, length(Y))
X ← cbind(X0, X)
```

a. Create a scatterplot matrix for X_1, X_2, X_3, X_4, X_5 and Y with the command

```
splom(~biomass[ ,3:8], aspect="fill")
```

Describe what this scatterplot matrix reveals about relationships between the variables.

In examining a plot for possible trends or patterns, it is sometimes helpful to pass smooth

curves through the points on the plot. The following code creates a function that inserts points and a smooth curve on a plot. Then it is applied to the last 6 columns of the biomass data frame with the pairs( ) function. The par( ) function sets the size and other features of the plot, such as thickness of lines and type and size of plotting symbols.

```
points.lines <- function(x, y)
   {
    points(x, y)
    lines(loess.smooth(x, y, 0.90))
   }

par(din=c(7,7), pch=18, mkh=.15, cex=1.2, lwd=3)
pairs(biomass[ ,-(1:2)], panel=points.lines)
```

You can compute a correlation matrix with the following code. The cor( ) function computes the correlations and the round( ) function rounds the printed correlations to 4 digits after the decimal point.

```
round(cor(biomass[-(1:2)]),4)
```

b. Use the qr( ) function to compute the rank of X.

c. Compute a vector of estimated regression coefficients

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}$$

Use matrix operations to do this. Do not use any built in S-PLUS functions for model fitting.

d. Compute the vector of estimated means, $\hat{\mathbf{Y}} = X\mathbf{b}$, and the vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Plot the residuals against the estimated means. This can be done with the following code. Some information on the par() function is included as comment statements.

```
# Specify plotting symbol and size of graph in inches.
# fin=c(w,h) specifies a plot that is w inches wide
#           and h inches high
# pch=18  requests a filled diamond as a plotting
#          symbol
# mkh=b   requests plotting symbols that are  b
#          inches high
# cex=c   sets the size of the characters used to
#          print labels at c times the printer default
# mar     mar=c(5,5,4,2) specifies the number of lines
#          of text on each side of the plot, starting
#          at the bottom and moving clockwise.  Here
#          write up to 5 lines at the bottom, up to
#          5 lines on the left side, etc…
```

```
par(fin=c(8.0,8.0),pch=18,mkh=.1,cex=1.5,mar=c(5,5,4,2))

b <- solve(t(X)%*%X)%*%t(X)%*%Y
yhat<-X%*%b
e <- Y-yhat

plot(yhat,e,xlab="Predicted Values",ylab="Residuals",
   main="Residual Plot")
```

This code stores the residuals in the vector `e` and the predicted values in the vector `yhat`. What does the plot indicate?

e. The residuals can be plotted against salinity with the following code:

```
plot(biomass$salinity,e,
      xlab="Salinity",ylab="Residuals",main="Residual Plot")
lines(loess.smooth(biomass$salinity, e, 0.90))
```

Plot the residual against each of the explanatory variables. What do these plots suggest?

f. Create a normal probability plot from the values in the residual vector with the following code:

```
qqnorm(e, main="Normal Probability Plot")
qqline(e)
```

What does this plot suggest?

g. Compute the sum of squared residuals and the corresponding residual mean square, which we will call s^2.

h. Sometimes you may want to write the contents of a matrix out to a file. This can be done with the following function (from Venables & Ripley). This function puts the column names on the first line of the output file.

```
write.matrix <- function(x, file = "", sep = " "){
     x <- as.matrix(x)
     p <- ncol(x)
     cat(dimnames(x)[[2]], format(t(x)), file = file,
           sep = c(rep(sep, p - 1), "\n"))
   }
```

Enter this function into the S-PLUS command window. The following code collects the results of the regression analysis into a matrix that includes columns for the case numbers, the explanatory variables, $\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\mathbf{e}$. The round( ) function is used to print The matrix in the command window with all entries rounded to 4 digits after the decimal point.

```
case<-1:45
heading <- c("Case","Salinity", "pH", "K", "Na", "Zn",
    "Biomass", "Predicted", "Residuals")
temp <-  cbind(case, X[ ,-1], Y, yhat, e)
dimnames(temp) <- list(case, heading)
round(temp,4)
```

Now the write.matrix( ) function is used to write this matrix to the file "c:/stat511/temp.out". You should change to file name to indicate where you want to save the file on your computer.

```
write.matrix(temp,file="c:/courses/st511/hw/temp.out")
```

Compute the estimate of the covariance matrix for **b**. The formula for this matrix is $s^2(X^TX)^{-1}$. Use this result to obtain standard errors for the estimated regression coefficients. Do not report the value of $s^2(X^TX)^{-1}$. Modify the S-PLUS code to create a matrix with the estimated coefficients in the first column and the corresponding standard errors in the second column. Label the rows and columns of your matrix. Use the write.matrix( ) function to write it out to a file. Submit a listing of the matrix and your S-PLUS code.

i. In this problem we entered the data into S-PLUS as a data frame, and converted parts of the data frame to matrices. Describe the difference between a matrix and a data frame in S-PLUS?