**Note:** Part I is a kind of review problems about the materials you already know and Part II covers estimable functions and ANOVA.

**Due:** Monday, April 9, in class

## PART I

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 $B = [1, 0, 0, 0]$ is a generalized inverse for $A$.

   (b) Find two other generalized inverses for $A$.

3. Let $X = (x_1, x_2, \cdots, x_n)^T$ 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 $n \times n$ 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, $\Delta_r$ is an $r \times r$ diagonal matrix containing the positive eigenvalues of $A$, and $L$ is an $n \times n$ orthogonal matrix where the columns are eigenvectors of $A$. Show that

$$G = L \begin{bmatrix} \Delta_r^{-1} & 0 \\ 0 & 0 \end{bmatrix} L^T$$

   satisfies the definition of the Moore- Penrose inverse of $A$.

5. (a) Use the **eigen( )** function in R to compute the eigenvalues and eigenvectors of

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

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

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

$$\sum_{i=1}^{n}(Y_i - b_1 X_{1i} - b_2 X_{2i} - \cdots - b_k X_{ki})^2 = (\mathbf{Y} - X\mathbf{b})^T (\mathbf{Y} - X\mathbf{b}).$$

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

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

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

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

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

7. In this problem we will use line commands to enter data into R 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 blackboard. 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 |
| $X_2$ | = 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 504 directory. Start R and get into the Command window. You can enter these data into R 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 $\mathbf{Y}$ 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

```
pairs(biomass[ ,3:8])
```

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(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 R 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.
# 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(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) 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)
```

Using the matrix obtained, compute the estimate of the covariance matrix for $\mathbf{b}$. The formula for this matrix is $s^2(X^TX)^{-1}$. Use this result to obtain standard errors for the estimated regression coefficients.

**PART II**

1. A food scientist performed the following experiment to study the effects of combining two 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 |
| | 6.7 | 7.1 | 5.5 |
| Fat 1 | 4.3 | 5.9 | 6.4 |
| | 5.7 | 5.6 | 5.8 |
| | 5.9 | 5.6 | 6.4 |
| Fat 2 | 7.4 | 6.8 | 5.1 |
| | 7.1 | | 6.2 |
| | | | 6.3 |

Consider the model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim NID(0, \sigma^2)$ 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) $\mu$

(b) $\alpha_2$

(c) $\beta_2 - \beta_3$

(d) $\gamma_{23}$

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

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

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

(h) $(\beta_2 - \beta_3) + 0.5(\gamma_{12} + \gamma_{22} - \gamma_{13} - \gamma_{23})$

(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 | $Y$: Yield(grams) | $T$ : Temperature(C) | 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

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

where

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

$\alpha_i$ 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 $E(\mathbf{Y}) = X\boldsymbol{\gamma}$. Write out the model matrix $X$ corresponding to the parameter vector $\boldsymbol{\gamma} = (\mu,\ \alpha_1,\ \alpha_2,\ \beta)^T$.

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

i. $\mu$

ii. $\mu + \alpha_2$

iii. $\beta$

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 **hw4p2.txt** on the blackboard. 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:/stat504/hw4p2.txt",header= T)
```

to enter these data into a data frame in R. Of course, you should replace "c:/stat504/hw4p2.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). For deriving the generalized inverse of $X^TX$, call in the "MASS" library. 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)
```

R will looking the MASS library for the ginv( ) function.

(d) Use R 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 $X^TX\mathbf{b} = X^T\mathbf{Y}$. Report your value of $\mathbf{b}$.

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

```
# R code for problem 2 on hw4, 2010
# Enter the data into a data frame
  W <- read.table("c:/stat504/hw4p2.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(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
  points(temp[1:5],Y[1:5],pch=1)
  points(temp[6:10],Y[6:10],pch=16)
  lines(temp[1:5],lm.out$fitted[1:5],lty=1)
  lines(temp[6:10],lm.out$fitted[6:10],lty=6)

   }
```

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

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

and

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

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

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

$$\hat{\mu} = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{Y} = \bar{Y}_{\cdot\cdot},$$

the sample mean of the ten observations.

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

(c) Give a formula for the vector of estimated means, $\hat{\mathbf{Y}} = P_{\mathbf{1}}\mathbf{Y}$, as a function of $\bar{Y}_{..}$.

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

$$SS_{total,uncorrected} = \mathbf{Y}^T\mathbf{Y} = \mathbf{Y}^T P_{\mathbf{1}}\mathbf{Y} + \mathbf{Y}^T(I - P_{\mathbf{1}})\mathbf{Y} = SS_{model,uncorrected} + SS_{residuals}$$

(e) For the "common mean" model,

$$SS_{residuals} = \mathbf{Y}^T(I - P_{\mathbf{1}})\mathbf{Y} = \sum_{i=1}^{2}\sum_{j=1}^{5}(Y_{ij} - \bar{Y}_{..})^2$$

is the quantity that one generally calls the corrected total sum of squares. Obtain a formula for

$$SS_{model,uncorrected} = \mathbf{Y}^T P_{\mathbf{1}}\mathbf{Y}$$

as a function of $\bar{Y}_{..}$.

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

$$SS_{total,uncorrected} = \mathbf{Y}^T\mathbf{Y} = \mathbf{Y}^T P_{\mathbf{1}}\mathbf{Y} + \mathbf{Y}^T(P_X - P_{\mathbf{1}})\mathbf{Y} + \mathbf{Y}^T(I - P_X)\mathbf{Y}$$

where $P_X$ 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

$$SS_{model,problem2} = \mathbf{Y}^T(P_X - P_{\mathbf{1}})\mathbf{Y} = SS_{residuals,commonmeansmodel} - SS_{residuals,problem2}$$

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)=2d.f..