STAT 8004 – Statistical Methods II
Spring 2015
Homework Assignment 4 – Solutions

1. In the context of Problems 2 of Homework Assignment 3, use R matrix calculations to do the following in the (non-full-rank) Gauss-Markov normal linear model

   (a) Find 90% two-sided confidence limits for σ.

   (b) Find 90% two-sided confidence limits for $\mu + \tau_2$.

   (c) Find 90% two-sided confidence limits for $\tau_1 - \tau_2$.

   (d) Find a p-value for testing the null hypothesis $H_0 : \tau_1 - \tau_2 = 0$ vs H_a : not H_0.

   (e) Find 90% two-sided prediction limits for the sample mean of $n = 10$ future observations from the first set of conditions.

   (f) Find 90% two-sided prediction limits for the difference between a pair of future values, one from the first set of conditions (i.e. with mean $\mu+\tau_1$) and one from the second set of conditions (i.e. with mean $\mu + \tau_2$).

   (g) Find a p-value for testing $H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$. What is the practical interpretation of this test?

   (h) Find a p-value for testing $H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \end{pmatrix}$.

Solutions:

```
library(MASS)
X <- matrix(c(rep(1,8),rep(c(rep(0,6),1),3),1),6,5)
Y <- c(2,1,4,6,3,5)
alpha <- 0.10

#a)
b <- ginv(t(X)%*%X,tol=1e-10)%*%t(X)%*%Y
df <- length(Y)-qr(X)$rank
SSE <- t(Y-X%*%b)%*%(Y-X%*%b)
ll <- sqrt(SSE/qchisq(1-alpha/2,df))
ul <- sqrt(SSE/qchisq(alpha/2,df))
```

```
c(ll,ul)
> c(ll,ul)
[1] 0.6459568 4.9365633

#b)
cvector <- c(1,0,1,0,0)
MSE <- SSE/df
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*cXXc)
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] 0.7353569 7.2646431

#c)
cvector=c(0,1,-1,0,0)
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*cXXc)
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] -6.498355  1.498355

#d)
t.ratio <- (cvector%*%b)/se
p.value <- 2*(1-pt(abs(t.ratio),df))
> p.value
          [,1]
[1,] 0.2094306

#e)

cvector <- c(1,1,0,0,0)
n <- 10; gamma <- 1/n
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*(gamma+cXXc))
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] -1.028782  4.028782
```

```
#f)
cvector <- c(1,1,0,0,0) - c(1,0,1,0,0)
gamma <- 2
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*(gamma+cXXc))
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] -8.607588  3.607588

#g) We are testing the equality of the treatment effect.
C <- matrix(c(0,1,-1,0,0,0,1,0,-1,0,0,1,0,0,-1),3,5,byrow=T)
d <- c(0,0,0)
df1 <- dim(C)[1]
CXXC <- C%*%ginv(t(X)%*%X,tol=1e-10)%*%t(C)
SSH0 <- t(C%*%b-d)%*%solve(CXXC)%*%(C%*%b-d)
F.ratio <- (SSH0/df1)/MSE
p.value <- 1-pf(F.ratio,df1,df)
> p.value
          [,1]
[1,] 0.2064399

#h)
C <- matrix(c(0,1,-1,0,0,0,0,1,-1,0),2,5,byrow=T)
d <- c(10,0)
df1 <- dim(C)[1]
CXXC <- C%*%ginv(t(X)%*%X,tol=1e-10)%*%t(C)
SSH0 <- t(C%*%b-d)%*%solve(CXXC)%*%(C%*%b-d)
F.ratio <- (SSH0/df1)/MSE
p.value <- 1-pf(F.ratio,df1,df)
> p.value
           [,1]
[1,] 0.01338688
```

2. In the following, make use of the data in Problem 4 of Homework Assignment 3. Consider a regression of y on $x_1, x_2, \ldots, x_5$. Use R matrix calculation to do the following in a full rank Gauss-Markov normal linear model.

   (a) Find 90% two-sided confidence limits for σ.

   (b) Find 90% two-sided confidence limits for the mean response under the conditions of data point #1.

   (c) Find 90% two-sided confidence limits for the difference in mean responses under the conditions of data points #1 and #2.

   (d) Find a p-value for testing the hypothesis that the conditions of data points #1 and #2 produce the same mean response.

   (e) Find 90% two-sided prediction limits for an additional response for the set of conditions $x_1 = 0.005$, $x_2 = 0.45$, $x_3 = 7$, $x_4 = 45$, and $x_5 = 6$.

   (f) Find a p-value for testing the hypothesis that a model including only x_1, x_3 and x_5 is adequate for "explaining" home price. (Hint: write it in the form of $H_0 : \mathbf{C}\boldsymbol{\beta} = 0$).

Solutions:

```
#a)
b <- ginv(t(X)%*%X,tol=1e-10)%*%t(X)%*%Y
df <- length(Y)-qr(X)$rank
SSE <- t(Y-X%*%b)%*%(Y-X%*%b)
ll <- sqrt(SSE/qchisq(1-alpha/2,df))
ul <- sqrt(SSE/qchisq(alpha/2,df))
> c(ll,ul)
[1] 5.610624 6.226291

#b)
cvector=X[1,]
MSE <- SSE/df
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*cXXc)
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] 25.21142 26.19733

#c)
cvector=X[1,]-X[2,]
MSE <- SSE/df
```

```
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*cXXc)
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] 1.202479 2.612541

#d)
t.ratio <- (cvector%*%b)/se
p.value <- 2*(1-pt(abs(t.ratio),df))
> p.value
             [,1]
[1,] 1.019758e-05



#e)
cvector=c(0.005,0.45,7,45,6)
gamma <- 1
cXXc <- cvector%*%ginv(t(X)%*%X,tol=1e-10)%*%cvector
se <- sqrt(MSE*(gamma+cXXc))
ll <- cvector%*%b - qt(1-alpha/2,df)*se
ul <- cvector%*%b + qt(1-alpha/2,df)*se
> c(ll,ul)
[1] 19.90023 39.40286

#f)
C <- matrix(c(0,0,1,0,0,0,0,0,0,0,1,0),2,6,byrow=T)
d <- c(0,0)
df1 <- dim(C)[1]
CXXC <- C%*%ginv(t(X)%*%X,tol=1e-10)%*%t(C)
SSH0 <- t(C%*%b-d)%*%solve(CXXC)%*%(C%*%b-d)
F.ratio <- (SSH0/df1)/MSE
p.value <- 1-pf(F.ratio,df1,df)
> p.value
             [,1]
[1,] 3.190781e-13
```

3. (a) In the context of Problem 1, part g), suppose that in fact $\tau_1 = \tau_2, \tau_3 = \tau_4 = \tau_1 - d\sigma$ What is the distribution of the F statistic?

solution The numerator has a non-central χ_2 distribution with 3 degrees of freedom and non-centrality parameter $(3/2)d^2$. The truth implies that

$$\mathbf{C}\boldsymbol{\beta} - d = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ d\sigma \\ d\sigma \end{pmatrix}.$$

Note that $(\mathbf{C}(\mathbf{X}^T\mathbf{X})\mathbf{C}^T)^{-1} = \begin{pmatrix} 5/6 & -1/6 & -1/3 \\ -1/6 & 5/6 & -1/3 \\ -1/3 & -1/2 & 4/3 \end{pmatrix}$, thus the noncentrality parameter is

$$\frac{1}{\sigma^2}(\mathbf{C}\boldsymbol{\beta} - d)^T(\mathbf{C}(\mathbf{X}^T\mathbf{X})\mathbf{C}^T)^{-1}(\mathbf{C}\boldsymbol{\beta} - d) = (3/2)d^2.$$

The denominator is independent of the numerator and has a central χ^2 distribution with 2 degrees of freedom. Therefore, the F-statistic has a non-central F distribution with (3,2) degrees of freedom and non-centrality parameter $(3/2)d^2$.

(b) Use R to plot the power of an $\alpha = 0.05$ level test as a function of d for $d \in [-5, 5]$, that is plotting $P(F >$ the cut-off value) against d. The R function `pf(q,df1,df2,ncp)` will compute cumulative (non-central) F probabilities for you corresponding to the value `q`, for degrees of freedom `df1` and `df2` when the noncentrality parameter is `ncp`.

Solutions

```
d <- seq(-5,5,length=100)
plot(d,1-pf(qf(0.95,3,2),3,2,(3/2)*d^2),type="l",xlab='d',ylab='power')
```