STAT 8004, Assignment 4

David Dobor

February 19, 2015

Question 1

In the context of Problem 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 $H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$. What is the practical interpretation of this test?

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

Answer to Question 1

The context is the one-way ANOVA Gauss-Markov model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ for the jth individual of the ith group (4 groups with sample sizes 2, 1, 1, 2 for groups, respectively) as follows:

$$\begin{bmatrix} 2 \\ 1 \\ 4 \\ 6 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{41} \\ \epsilon_{42} \end{bmatrix}$$

(a) With $n = 6$ observations and the design matrix being of rank 4, we find that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-\text{rank(X)}} = \chi^2_2.$$

That is

$$P\left(\frac{\text{SSE}}{\text{upper 0.05 qt of } \chi^2_2} < \sigma^2 < \frac{\text{SSE}}{\text{lower 0.05 qt of } \chi^2_2} \right) = 0.9.$$

```r
# compute sum of squared errors
beta.hat <- ginv(t(X) %*% X) %*% t(X) %*% Y
Y.hat <- X %*% beta.hat;
SSE <- t(Y - Y.hat) %*% (Y - Y.hat) # ans: 2.5

# compute the endpoints for the 90% confidence interval
lower.limit <- SSE / qchisq(0.95, 2)  # ans: 0.4172603
upper.limit <- SSE / qchisq(0.05, 2)  # ans: 24.36966

c(sqrt(lower.limit), sqrt(upper.limit))
#ans: 0.6459568 4.9365633
```

The 90% confidence interval for σ is given by: $\boxed{(0.6459\,,\ 4.9366)}$

(b) Here $c^T = (1,0,1,0,0)$ (we have $c^T\beta = \mu+\tau_2$). We note that $c^T\beta$ is an estimable function (c^T is the third row of X) and compute the two sided 90% confidence interval as follows:

```
c <- matrix(c(1, 0, 1, 0, 0), 5, 1)
c.beta.hat <- t(c) %*% beta.hat #= 4
MSE <- SSE / df

# 90% two sided confidence interval
c.beta.hat +
    c(-1, 1) * qt(.95, df)* sqrt(MSE) * sqrt(t(c) %*% XtXi %*% c)
#ans: 0.7353569 7.2646431
```

The 90% confidence interval for $\mu+\tau_2$ is given by: $\boxed{(0.7353569\ ,\ 7.2646431)}$

(c) Here $c^T = (0,1,-1,0,0)$ (we have $c^T\beta = \tau_1-\tau_2$). We note that $c^T\beta$ is an estimable function (c^T is (row 2 - row 3) of X) and compute the two sided 90% confidence interval as follows:

```
c <- matrix(c(0, 1, -1, 0, 0), 5, 1)
c.beta.hat <- t(c) %*% beta.hat #= -2.5

# 90% two sided confidence interval
c.beta.hat +
    c(-1, 1) * qt(.95, df) * sqrt(MSE) * sqrt(t(c) %*% XtXi %*% c)
#ans: -6.498355  1.498355
```

The 90% confidence interval for $\tau_1-\tau_2$ is given by: $\boxed{(-6.498355\ ,\ 1.498355)}$

(d) Here

$$H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} = 0$$

And we compute the following F ratio

$$F = \frac{\text{SSH}_0\ /1}{\text{SSE}\ /\ 2} = \frac{\text{MSH}_0}{\text{MSE}}$$

as follows:

```r
# here c and c.beta.hat are the same as in part (c)
# sum of squares under the null (numerator in the F test):
SSH <-
    t(c.beta.hat) %*% ginv( (t(c) %*% XtXi %*% c) ) %*% c.beta.hat

SSE <- t(Y - Y.hat) %*% (Y - Y.hat)
MSH <- SSH
MSE <- SSE / df

# the F ratio
F <- MSH / MSE
# the p-value
1 - pf(F, 1, 2) #ans: 0.2094306

### alternatively, we could use this:
t.stat <- c.beta.hat / (sqrt(MSE) * sqrt(t(c) %*% XtXi %*% c))
p.value <- 2*(1 - pt(abs(t.stat),df))
###
# the p.value is the same as with the F-test
###
```

The p-value here is $\boxed{0.2094306}$

(e) Following the notation used in class, for 10 future observations from the first set of conditions we set c^T to be the first row of X: $\mathbf{c}^T = (1,1,0,0,0)$ (thus $\mathbf{c}^T\beta$ is clearly estimable), and set $\gamma = 1/10$. Then $\text{var}(y^*) = 1/10$.

Thus
$$\widehat{\mathbf{c}^T\beta} - y^* \sim N\left(0, \sigma^2(\gamma + \sigma^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^-\mathbf{c})\right)$$

independently of SSE.

The t-test is then as follows:

```r
c <- matrix(c(1, 1, 0, 0, 0), 5, 1)
c.beta.hat <- t(c) %*% beta.hat #= 1.5
gamma <- 1/10
MSE <- SSE / df

c.beta.hat + c(-1,1) * qt(.95, df) * sqrt(MSE) *sqrt(gamma + t(c)
   %*% XtXi %*% c)

# ans: -1.028782  4.028782
```

Thus the 90% two-sided prediction limits for the sample mean of 10 future observations from the first set of conditions are $\boxed{(-1.028782, 4.028782)}$.

(f) For the difference of 2 future values we set $\gamma = 2$ and $\mathbf{c}^T = (2,1,1,0,0)$. The rest is similar to part (e), as follows:

```
c <- matrix(c(2, 1, 1, 0, 0), 5, 1)
c.beta.hat <- t(c) %*% beta.hat #= 5.5
gamma <- 2
MSE <- SSE / df

c.beta.hat + c(-1,1) * qt(.95, df) * sqrt(MSE) *sqrt(gamma + t(c)
    %*% XtXi %*% c)
# ans: -0.607588 11.607588
```

Thus the 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 $\mu + \tau_2$) are $\boxed{(-0.607588, 11.607588)}$.

(g) The practical interpretation here is that the effects for groups 2, 3, and 4 (the values τ_1, τ_2, τ_3) are not that different from the effect for group 1 (from the value of τ_1).

Parts (g) and (h) are similar to (d). Results follow in this R code:

```
C <- t(matrix(c(0, 1, -1, 0, 0,
                0, 1, 0, -1, 0,
                0, 1, 0, 0, -1), nrow=5, ncol=3))
C.beta.hat <- C %*% beta.hat
# sum of squares under the null (numerator in the F test):
SSH <-
    t(C.beta.hat) %*% ginv( (C %*% XtXi %*% t(C)) ) %*% C.beta.hat
MSH <- SSH / 3

SSE <- t(Y - Y.hat) %*% (Y - Y.hat)
MSE <- SSE / df
# the F ratio
F <- MSH / MSE
1 - pf(F, 1, 2) #ans: 0.1835034
```

The p-value is $\boxed{0.1835034}$

(h)
```r
C <- t(matrix(c(0, 1, -1, 0, 0,
                0, 0, 1, -1, 0), nrow=5, ncol=2))
d <- matrix(c(10,0))
u <- C %*% beta.hat - d
# sum of squares under the null (numerator in the F test):
SSH <-
    t(u) %*% ginv( (C %*% XtXi %*% t(C)) ) %*% u
SSE <- t(Y - Y.hat) %*% (Y - Y.hat)
MSH <- SSH /2
MSE <- SSE / df
# the F ratio
F <- MSH / MSE
1 - pf(F, 1, 2) #ans: 0.01329846
```

The p-value is $\boxed{0.01329846}$

Question 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 $H_0 : \mathbf{C}\beta = 0$).

Answer to Question 2

(a) The `Boston` dataset contains $n = 506$ observations. Also, $\text{rank}(X) = 6$. So

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-\text{rank(X)}} = \chi^2_{500}.$$

That is

$$P\left(\frac{\text{SSE}}{\text{upper 0.05 qt of } \chi^2_{500}} < \sigma^2 < \frac{\text{SSE}}{\text{lower 0.05 qt of } \chi^2_{500}}\right) = 0.9.$$

```
# after loading the data as in assignment 3, we do:
# compute sum of squared errors
beta.hat <- ginv(t(X) %*% X) %*% t(X) %*% Y
Y.hat <- X %*% beta.hat;
SSE <- t(Y - Y.hat) %*% (Y - Y.hat) # ans: 17411.94

# compute the endpoints for the 90% confidence interval
lower.limit <- SSE / qchisq(0.95, 500)  # ans: 31.4791
upper.limit <- SSE / qchisq(0.05, 500)  # ans: 38.7667
```

Thus the 90% confidence interval for σ is:

$$\left(\sqrt{31.4791},\ \sqrt{38.7667}\right)$$

$$\boxed{(5.610624,\ 6.226291)}$$

(b)

```
c <- X[1,]    # data point # 1: first row
c.beta.hat <- t(c) %*% beta.hat # ans: 25.70437

# 90% two sided confidence interval
c.beta.hat +
    c(-1, 1) * qt(.95, df)* sqrt(MSE) * sqrt(t(c) %*% XtXi %*% c)

# ans: 25.21142 26.19733
```

Answer to (b):

$$\boxed{(25.21142,\ 26.19733)}$$

(c)

```
c <- X[1,] - X[2,]    # data point # 1: first row - second row
c.beta.hat <- t(c) %*% beta.hat #

# 90% two sided confidence interval
c.beta.hat +
    c(-1, 1) * qt(.95, df)* sqrt(MSE) * sqrt(t(c) %*% XtXi %*% c)

# ans: 1.202479 2.612541
```

Answer to (c):

$$\boxed{(1.202479,\ 2.612541)}$$

(d)

```
# here c and c.beta.hat are the same as in part (c)
c <- X[1,] - X[2,]    # data point # 1: first row - second row
c.beta.hat <- t(c) %*% beta.hat

# two sided t-test
p.value <- 2*(1 - pt(abs(c.beta.hat / (sqrt(MSE) * sqrt(t(c) %*%
   XtXi %*% c))),df))
# ans: 1.019758e-05
```

Answer to (d):

$$\boxed{0.00001019758}$$

(e)
```
c <- c(1, 0.005, 0.45, 7, 45, 6)
se <- sqrt(MSE)*sqrt(1 + c %*% XtXi %*% c)   #ans: 5.917394
c.beta.hat <- t(c) %*% beta.hat

c.beta.hat + c(-1, 1) * qt(.95, df) * se
# ans: 19.90023 39.40286
```

Answer to (e):

$$\boxed{(19.90023,\ 39.40286)}$$

(f) We interpret the hypothesis "a model including only x_1, x_3 and x_5 is adequate for explaining home price" as $H_0 : \beta_2 = \beta_4 = 0$ and write H_0 in the form $\mathbf{C}\beta = 0$:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We have that the following ratio is F distributed

$$F = \frac{\text{SSH}_0 \ / \ 2}{\text{SSE} \ / \ 500}$$

```
C <- matrix(
    c(0, 0,
      0, 0,
      1, 0,
      0, 0,
      0, 1,
      0, 0),
    nrow=2,
    ncol=6)

XtXi <- ginv(t(X) %*% X);
C.beta.hat <- C %*% beta.hat

# sum of squares under the null (numerator in the F test):
SSH <-
    t(C.beta.hat) %*% ginv( (C %*% XtXi %*% t(C)) )  %*% C.beta.
       hat

# squared errors (same as before, part (a) denominator in the F
   test):
```

```
SSE <- t(Y - Y.hat) %*% (Y - Y.hat) # ans: 17411.94
MSE <- SSE / df

# the F-ratio and the p-value
F <- (SSH / 2) / MSE
1 - pf(F, 2, 500)    #ans: 3.190781e-13
```

Thus the p-value is negligeable (3.190781e-13).
(this tiny p-value somehow doesn't sit well with me. Perhaps I've tested the wrong hypothesis)

Question 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?

(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 non-centrality parameter is `ncp`.

Answer to Question 3

Given that

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{bmatrix} = \begin{bmatrix} 0 \\ d\sigma \\ d\sigma \end{bmatrix}$$

We will have a non-central F distribution with 3 and 2 degrees of freedom (numerator of the F-ratio has a χ^2 distribution with 3 degrees of freedom; denominator has 2). We compute the non-centrality parameter based on the following quantities.

$$(\mathbf{C}(\mathbf{X}^T\mathbf{X})\mathbf{C}^T)^{-1} = \begin{bmatrix} 5/6 & -1/6 & -1/3 \\ -1/6 & 5/6 & -1/3 \\ -1/3 & -1/2 & 4/3 \end{bmatrix}$$

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

Then *(reading the textbook would make one think that using 1/2 of this paramter would be the way to go. However, I'm following the lecture notes here)*:

$$\boxed{\text{non-centrality parameter} = \frac{3}{2}d^2}$$

(a) The distribution is the non-central F with parameters $(3, 2, \frac{3}{2}d^2)$

(b) The following R code produces the graph shown below:

```
d <- seq(-5,5,.1)
power <- 1-pf(qf(0.95, 3, 2), 3, 2, (3/2)*d^2)

library(ggplot2)
qplot(d, power, geom="line",
      main="Power Calculations at the 0.05 Significance Level\n",
      xlab="paramter d")

ggsave(file="power.png")
```