STAT 8004, Assignment 4

David Dobor

February 18, 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

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

```
# 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 / (n - rank.X)
df <- n - rank.X

# 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:  -3.998355  3.998355
```

The 90% confidence interval for $\tau_1 - \tau_2$ is given by: $\boxed{(-3.998355\ ,\ 3.998355)}$

(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:

```
c.beta.hat <- c %*% beta.hat    #= -2.5

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

MSH <- SSH
MSE <- SSE / (n - rank.X)

# the F ratio
F <- MSH / MSE

1 - pf(F, 1, 2)   #ans: 0.2094306
```

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

(e)

(f)

(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 (τ_1).

Parts (g) and (h) are similar to (d) and we only present the R code and results.

```
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

SSE <- t(Y - Y.hat) %*% (Y - Y.hat)

MSH <- SSH
MSE <- SSE / (n - rank.X)

# the F ratio
F <- MSH / MSE

1 - pf(F, 1, 2)   #ans: 0.0741799
```

The p-value is $\boxed{0.0741799}$

(h) The p value obtained here is really small: $\boxed{0.006715993}$, and we are able to reject the null more comfortably. This is not surprising, considering that the first hypothesis says $\tau_1 - \tau_2 = 10$. One would hardly expect such difference in the τ_1 and τ_2 effects given the observed responses in the two groups.

The code follows:

```
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
MSE <- SSE / (n - rank.X)

# the F ratio
F <- MSH / MSE

1 - pf(F, 1, 2)   #ans: 0.006715993
```

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, 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:

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

(b)

(c)

(d)

(e)

(f) ——> COME BACK - RUSULTS LIKELY WRONG! <——

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 non-central 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 in part (a)):
SSE <- t(Y - Y.hat) %*% (Y - Y.hat) # ans: 17411.94
sigma.squared <- SSE / 500

# non cenrality parameter for F
ncp <- (1/2) * (1 / sigma.squared) * SSH

# the F-ratio and the p-value
F <- (SSH / 2) / (SSE/ 500)
1 - pf(F, 2, 500, ncp)    #ans: 0.01607116
```

Thus the p-value is 0.016.

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