8004 Homework 4

Nooreen Dabbish
February 19, 2015

1 Problem 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.

The model described in HW3, Problem 2 in $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ matrix form is:

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

Because the problem statement says this is a Gauss-Markov normal linear model, we know that $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\beta, \sigma^2\ \mathbf{I})$.

(a).1 SSE/σ^2

Using theorem 1 in the Appendix, we can show:

$$\frac{SSE}{\sigma^2} = \frac{(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})}{\sigma^2} \sim \chi^2_{n-\text{rank}(X)}$$

Rearranging to find confidence limits for σ gives:

$$P\left(\sqrt{\frac{SSE}{\text{upper}\,\alpha/2\,\text{quantile of}\,\chi^2_{\text{n}-\text{rank(X)}}}} < \sigma < \sqrt{\frac{SSE}{\text{upper}\,\alpha/2\,\text{quantile of}\,\chi^2_{\text{n}-\text{rank(X)}}}}\right) = 1 - \alpha$$

(a).2 Solution from R

Using the hand-written function `sigmacalc`, included in the appendix. The following two-sided 90% confidence limits for σ were obtained: `0.646` $< \sigma <$ `4.9366`.

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

Using the t-distribution describing the distribution of estimable function c'β, the handwritten R function `cbetacalc` included in the appendix, was used to caluclate confidence limits for this entity, where c' = (1, 0, 1, 0 , 0).

`0.7354` $< \mu + \tau_2 <$ `7.2646`

(b).1 Estimable functions c'β

For an estimable $\mathbf{c}'\beta$, we have:

$$\frac{\widehat{\mathbf{c}'\beta} - \mathbf{c}'\beta}{\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}}} \sim t_{n-\text{rank}(X)}$$

Note that $MSE = \frac{SSE}{n-\text{rank}(X)}$. Rearranging to find 1 - α confidence limits for $\mathbf{c}'\beta$, denoting t* = the upper $\alpha/2$ quantile of $\mathrm{t}_{\text{n}-\text{rank(X)}}$, we have:

$$P\left(\widehat{\mathbf{c}'\beta} - t^\star\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}} < \mathbf{c}'\beta < \widehat{\mathbf{c}'\beta} + t^\star\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^-\mathbf{C}}\right) = 1 - \alpha$$

(c) Find 90% two-sided confidence limits for τ_1 - τ_2.

Proceeding as in part b, here τ_1 - τ_2 = $\mathbf{c}'\beta = (0,1,-1,0,0)\begin{pmatrix}\mu\\ \tau_1\\ \tau_2\\ \tau_3\\ \tau_4\end{pmatrix}$. The function `cbetacalc` was used once again with **c** above.

`-6.4984` $< \tau_1 - \tau_2 <$ `1.4984`

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

(d).1 General Linear Hypothesis Test

The general linear hypothesis test is the following F test for $H_0 : \mathbf{C}\beta = \mathbf{0}$ verus $H_1 : \mathbf{C}\beta \neq \mathbf{0}$, given $\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\beta, \sigma^2\,\mathbf{I})$, **C** q x (/k/+1), rank(**C**) = q, with SSH = the sum of squares due to the hypothesis or due to $\mathbf{C}\beta$. Note that

$\frac{SSH}{\sigma^2} = \frac{(\mathbf{C}\hat{\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\beta}}{\sigma^2} \sim \chi^2(q, \frac{(\mathbf{C}\beta)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\beta}{2\sigma^2})$

and

$\frac{SSE}{\sigma^2} = \frac{\mathbf{y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{\sigma^2} \sim \chi^2(n - rank(X))$.

Taking the ratio gives us our test statistic:

$$F = \frac{SSH/q}{SSE/(n - rank(X))}$$

- If $H_0 : \mathbf{C}\beta = \mathbf{0}$ is false, F ~ F(q,n-rank(X),λ), where $\lambda = \frac{(\mathbf{C}\beta)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\beta}{2\sigma^2}$).
- Notice that if $\mathbf{C}\beta = \mathbf{0}$ is true, λ defined above = 0, giving F ~ F(q, n-rank(X)).

(d).2 *p*-value from the F statistic

We need to find the F statistic described above. Here **C** is **a**' from above, **a**'=(0,1,-1,0,0), and **C** is 1 x 5, rank 1.

We used the handwritten function `Cbetahatd` throughout for General Linear Hypothesis Testing. It is included in the appendix for your reference.

The *p*-value obtained was `0.209430584957905`.

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

(e).1 A t statistic for prediction

Consider future observation y_0, $y_0 = \mathbf{x}_0'\,\beta + \epsilon_0$ with $\hat{y}_0 = \mathbf{x_0'}\hat{\beta}$, where $\hat{y}_0$ is computed from n observations and y_0 is obtained independently. We find that $E(y_0 - \hat{y}_0) = 0$ and

$var(y_0 - \hat{y}_0) = var(\epsilon_0) + var(\mathbf{x_0'}\hat{\beta}) = \sigma^2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]$, where $\widehat{var(y - \hat{y})} = s^2 2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]$. Because of the independence of s^2 and y_0 and $\hat{y}_0$, we have the following t statistic:

$$t = \frac{y_0 - \hat{y}_0 - 0}{s\sqrt{1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}} \sim t(n - \mathrm{rank}(X))$$

Therefore,

$$P = \left[-t_{\alpha/2, n-rank(X)} \leq \frac{y_0 - \hat{y}_0 - 0}{s\sqrt{1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}} \leq t_{alpha/2, n-rank(X)}\right] = 1 - \alpha$$

Re-arranging in terms of $\mathbf{x_0'}\hat{\beta} = \hat{y}_0$ gives:

$$\mathbf{x_0'}\hat{\beta} \pm t_{\alpha/2,n-rank(X)} s \sqrt{1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}}.$$

(e).2 Predicitions for n observations from $\mu + \tau_1$

Using the preceeding theory and the handwritten R function, `predict`, which is included in the appendix. I ran a prediction fo n =10 from the first condition $\mathbf{x}_0$ = (1,1,0,0,0).

The 90% confidence limits obtained for the mean were `-1.0288` to `4.0288`.

(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 $\mu + \tau_2$).

Similar to part (e) above, here I used my `predict` function again, except an n of .5 in order to obtain a gamma of 2 and a $\mathbf{x}_0$ vector of the difference of the first two conditions:

(1,1,0,0,0) - (1,0,1,0,0) = (0,1,-1,0,0).

This gave 90 % prediction limits for the difference as follows: `-8.6076` to `3.6076`.

(g) Find a p-value for testing the following: What is the practical interpretation of this test?

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

The null hypothesis asked by this test is whether $\tau_1 = \tau_2 = \mathrm{tau}_3 = \mathrm{tau}_4$, if all these parameters are equal there would be no difference among the treatments. I performed the test using the General Linear Hypothesis Testing function described above, `Cbetahatd`, with the matrix above as C in C'β and the d-vector = (0,0,0).

```
G <- t(matrix(c(0,1,-1,0,0,
                          0,1,0,-1,0,
                          0,1,0,0,-1),nrow=3,ncol=5, byrow=TRUE))
 Cbetahatd(Y1,X1,G,c(0,0,0))
```

I obtained a p value of 0.20643991448067, indicating that it is unlikely that all of the parameters are equal.

(h) Find a p-value for testing:

$$H_0 : \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 0 \end{pmatrix}.$$

In this test, the null hypothesis asks whether τ_1 - τ_2 = 10 and $\tau_2 = \tau_3$. I tested this hypotheis as in question 1g), using the General Linear Hypothesis and the F-test implemented in my function `Cbetahatd`, note that the vector (10,0) was entered for the **d** vector.

```
H <- t(matrix(c(0, 1, -1, 0, 0, 0, 0, 1, -1, 0), nrow=2, ncol=5, byrow=T))

   Cbetahatd(Y1,X1,H,c(10,0))
```

A significant p-value of 0.0134 was obtained, suggesting that this hypothesis is acceptable.

2 Problem 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 calculations to do the following in a full rank Gauss-Markov normal linear model.

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

Calling our `sigmacalc` function on the Boston data set, we find 90% confidence limits for sigma of `5.6106` $< \sigma <$ `6.2263`.

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

To find these 90% confidence limits, we will use the $t_{n-\text{rank(X)}}$-distribution of $\frac{\widehat{c'\beta}-c'\beta}{\sqrt{MSE}\sqrt{c'(X'X)^{-}c}}$, where c' is the first row of our data set (data point #1).

Using the `cbetacalc` function to do this, as `cbetacalc(YB,XB, .1, XB[1,])` we find a 90% confidence interval for the mean response under the conditions of data point #1 of `25.2114` to `26.1973`.

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

To find these 90% confidence limits, we will use the t-distribution function of c'β again, where c' is the difference beteen the first row of our data set and the second row (data points #1 and #2).

Using the `cbetacalc` function to do this, as `cbetacalc(YB,XB, .1, (XB[1,]-XB[2,]))` we find `1.2025` to `2.6125` is a 90% confidence interval for the difference in mean responses under conditions 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.

An F-test was used to test the hypothesis that the product between the vector describing the differences between conditions 1 and 2 and beta is **0**. That is H_0 : c'β = **0**, where c' = XB[1,] - XB[2,]. This was done using my general linear hypothesis testing function: `Cbetahatd(YB,XB, (XB[1,]-XB[2,]))`. The *p*-value obtained was `1.01975837067947e-05`.

(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 $x_5 = 6$.

90 % prediction limits for an additional response from these conditions were obtained using the conditions as our c-vector in the `predict` function: `predict(YB,XB, .1, c(1,0.005,0.45,7,45,6), 1)` . The limits obtained were `19.9002` to `39.4029`.

(f) Find a *p*-value for testing the hypothesis that a model including only x_1, x_3, and x_5 is adequare for "explaining" home price.

We use an F-test implementation of the General Linear Hypothesis test using the `Cbetahatd` function described previously. We are testing the hypothesis that $\beta_2 = \beta_4 = 0$, with a $\mathbf{C} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$.

To investigate this solution, we also

```
CB <- t(matrix(c(0,0,1,0,0,0,0,0,0,0,1,0),nrow=2,ncol=6,byrow=T))
Cbetahatd(YB,XB,CB)
```

This gives a p-value of 3.1907809727727e-13, heavily supporting the reduced model.

(f).1 Full–Reduced model approach

We can create a p-value to test these models using an F statistic, constructed out of the ratio of the difference in regression sum of squares between the full (SSR_{full}) and reduced($\text{SSR}_{\text{reduced}}$) models and the sum of squared error (SSE). These quantities are independent and follow a non-central $\chi^2(h,\lambda)$ and central $\chi^2(n\text{-}k\text{-}1)$ respectively where n is the number of observations, k is the number of parameters in the full model, and h is the difference in the number of parameters between the full and reduced models. The non-centrality parameter λ can be written β_2**'[$\mathbf{X}_2$'$\mathbf{X}_2$ - $\mathbf{X}_2$'$\mathbf{X}_1$($\mathbf{X}_1$'$\mathbf{X}_1$)$^{-1}$$\mathbf{X}_1$'$\mathbf{X}_2$]**$\beta_2$**/2**$\sigma^2$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ form a partition of $\mathbf{X}$ such that we can write:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \epsilon = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$$

And the reduced model would be $\mathbf{y} = \mathbf{X}_1\beta_1{}^\star + \epsilon^\star$.

create reduced model design matric and X1$_B andestimatorbhat1_B X1_B < -XB[,-c(3,5)]bhat1_B < -ginv(t(X1_B)SSI$
$-t(bhat1_B)$
YhatB <- XBSSE$_B < -t(YB - YhatB)$
F$_2 f < -((SSR_B f - SSR_B r)/2)/(SSE_B/(length(YB) - qr(XB)$rank))
pf$_2 f < -pf(F_2 f, 2, (length(YB) - (qr(XB)$rank)), lower.tail=FALSE)

This strategy arrives at a very similar p-value: `3.19090353910838e-13`.

3 Problem 3

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

The F statistic for Problem 1, part g is given by $F = \frac{Q/s}{SSE/N-\text{rank}(X)} \sim F(s, N - \text{rank}(X), \lambda)$.

Where $Q = (\widehat{C'\beta} - d)'(C'(X'X)^{-}C)^{-1}(\widehat{C'\beta} - d)$ and $\lambda = \frac{1}{\sigma^2}(C'\beta - d)'(C'(X'X)^{-}C)^{-1}(C'\beta - d)$.

Therefore, if $\tau_1 = \tau_2$, and τ_3=τ_4=τ_1 - dσ, our non-centrality parameter will equal

$$\lambda = \frac{1}{\sigma^2}(0, d\sigma, d\sigma)(C'(X'X)^{-})C)^{-1}\begin{pmatrix}0\\ d\sigma\\ d\sigma\end{pmatrix}.$$

Evaluating for (C'(X'X)$^{-}$C)$^{-1}$ in R, we find:

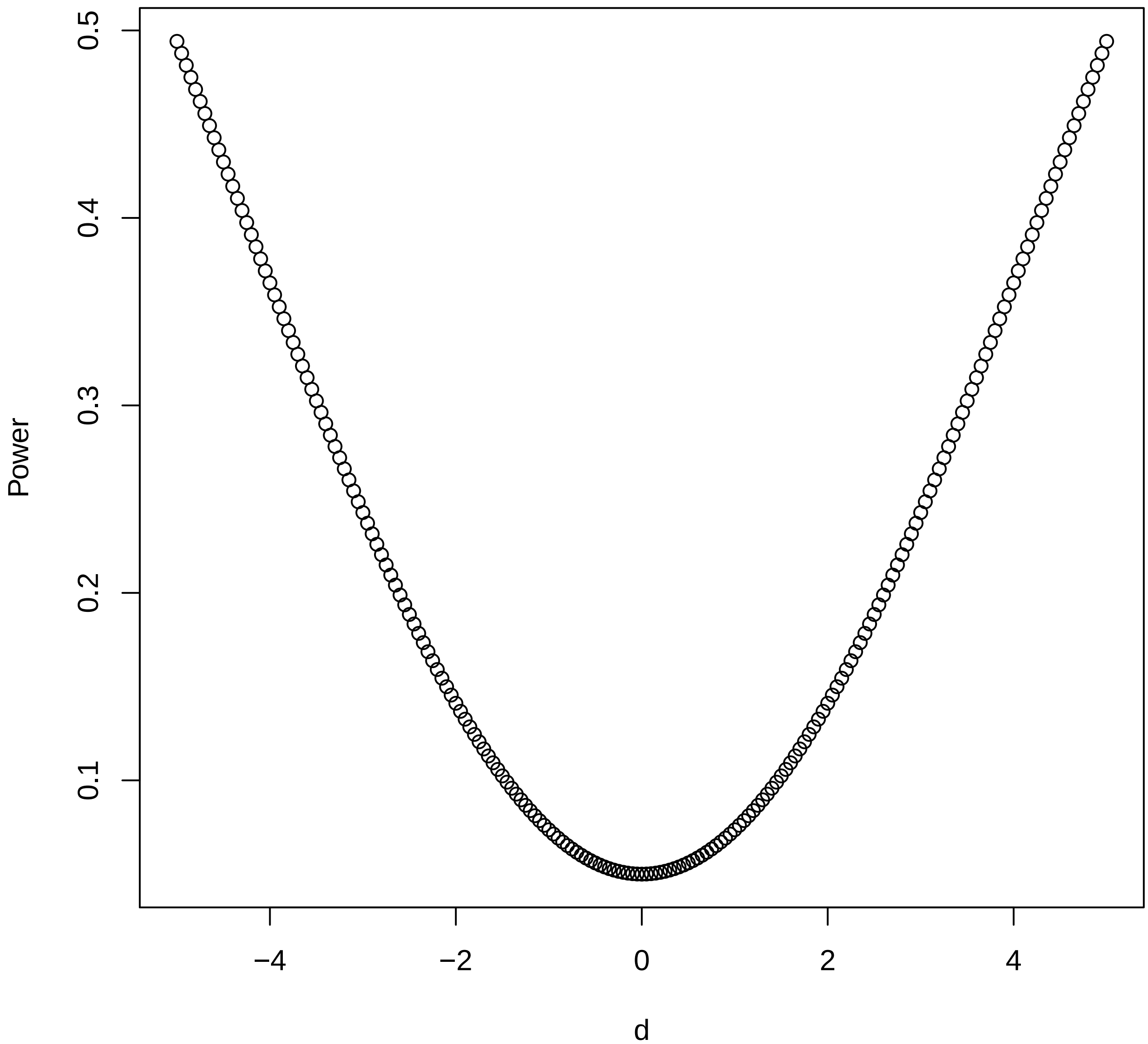
```
fractions(ginv(t(C1g)%*%ginv(t(X1)%*%X1)%*%C1g))
```

$$(C'(X'X)C)^{-1} = \begin{pmatrix} 5/6 & -1/6 & -1/3\\ -1/6 & 5/6 & -1/3\\ -1/3 & -1/3 & 4/3 \end{pmatrix}$$

Giving $\lambda = \frac{3}{2}d^2$ so the final distribution of the F statistic is F(3, 2, $\frac{3}{2}d^2$).

(b) Use R to plot the power of the α = 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.

```
d <- seq(-5,5,by=.05)
Power <- 1-pf(qf(0.95,3,2),3,2,1.5*d^2)
plot(d, Power)
```



r0.4 :

Figure 1: Power of an $\alpha = 0.05$ level test as a function of d.

4 Appendix: Additional Notes

(a) Useful Theorems

Theorem 4.1. *Suppose* $\mathbf{Y} \sim MVN_n(\mu, \mathbf{Sigma})$, Σ *positive definite. Also suppose* $\mathbf{A}_{n\times n}$ *symmetric and rank(*$\mathbf{A}$*)* $= k$. *If* $\mathbf{A}\Sigma$ *idempotent,* $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi^2_k(\mu'\mathbf{A}\mu)$.

Theorem 4.2. *Suppose* $\mathbf{Y} \sim MVN_n(\mu, \sigma^2\mathbf{I})$. *And the product* $\mathbf{BA} = \mathbf{0}$, *with A and B of appropriate size. Then,*

[(a)]If $\mathbf{A}$ *symmetric,* $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ *and* $\mathbf{BY}$ *are independent. If both* $\mathbf{B}$ *and* $\mathbf{A}$ *symmetric,* $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ *and* $\mathbf{Y}'\mathbf{B}\mathbf{Y}$ *are independent.*

(b) Distributions of interests

(b).1 SSE/σ^2

Using theorem 4.1 above, we can show:

$$\frac{SSE}{\sigma^2} = \frac{(\mathbf{Y}-\hat{\mathbf{Y}})'(\mathbf{Y}-\hat{\mathbf{Y}})}{\sigma^2} \sim \chi^2_{n-\text{rank}(X)}$$

Rearranging to find confidence limits for σ gives:

$$P\left(\sqrt{\frac{SSE}{\text{upper}\,\alpha/2\,\text{quantile of}\,\chi^2_{\text{n}-\text{rank(X)}}}} < \sigma < \sqrt{\frac{SSE}{\text{upper}\,\alpha/2\,\text{quantile of}\,\chi^2_{\text{n}-\text{rank(X)}}}}\right) = 1-\alpha$$

(b).2 Estimable functions c'β

For an estimable **c'**β, we have:

$$\frac{\widehat{\mathbf{c}'\beta} - \mathbf{c}'\beta}{\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}}} \sim t_{n-\text{rank}(X)}$$

Note that $MSE = \frac{SSE}{n-\text{rank}(X)}$. Rearranging to find 1 - α confidence limits for **c'**β, denoting $\text{t}^\star$ = the upper $\alpha/2$ quantile of $\text{t}_{\text{n}-\text{rank(X)}}$, we have:

$$P\left(\widehat{\mathbf{c}'\beta} - t^\star\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}} < \mathbf{c}'\beta < \widehat{\mathbf{c}'\beta} + t^\star\sqrt{MSE}\sqrt{\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}}\right) = 1-\alpha$$

(b).3 A t statistic for prediction

Consider future observation y_0, $\text{y}_0 = \mathbf{x}_0'\ \beta + \epsilon_0$ with $\hat{y}_0 = \mathbf{x_0'}\hat{\beta}$, where $\hat{y}_0$ is computed from n observations and y_0 is obtained independently. We find that $E(y_0 - \hat{y}_0) = 0$ and

$var(y_0 - \hat{y}_0) = var(\epsilon_0) + var(\mathbf{x_0'}\hat{\beta}) = \sigma^2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]$, where $\widehat{var(y-\hat{\ }y)} = s^2 2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]$. Because of the independence of s^2 and y_0 and $\hat{y}_0$, we have the following t statistic:

$$t = \frac{y_0 - \hat{y}_0 - 0}{s\sqrt{1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}} \sim t(n - \text{rank}(X))$$

Therefore,

$$P = \left[-t_{\alpha/2,n-rank(X)} \leq \frac{y_0 - \hat{y}_0 - 0}{s\sqrt{1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}} \leq t_{alpha/2,n-rank(X)}\right] = 1-\alpha$$

Re-arranging in terms of $\mathbf{x_0'}\hat{\beta} = \hat{y}_0$ gives:

$$\mathbf{x_0'}\hat{\beta} \pm t_{\alpha/2,n-rank(X)}\, s\sqrt{1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}.$$

(c) General Linear Hypothesis Test

The general linear hypothesis test is the following F test for $H_0 : \mathbf{C}\beta = \mathbf{0}$ verus $H_1 : \mathbf{C}\beta \neq \mathbf{0}$, given $\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I})$, $\mathbf{C}$ q x (/k/+1), rank($\mathbf{C}$) = q, with SSH = the sum of squares due to the hypothesis or due to $\mathbf{C}\beta$. Note that

$\frac{SSH}{\sigma^2} = \frac{(\mathbf{C}\hat{\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\hat{\beta}}{\sigma^2} \sim \chi^2(q, \frac{(\mathbf{C}\beta)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\beta}{2\sigma^2})$

and

$\frac{SSE}{\sigma^2} = \frac{\mathbf{y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{\sigma^2} \sim \chi^2(n - rank(X))$.

Taking the ratio gives us our test statistic:

$$F = \frac{SSH/q}{SSE/(n - rank(X))}$$

- If $H_0 : \mathbf{C}\beta = \mathbf{0}$ is false, F ~ F(q,n-rank(X),λ), where $\lambda = \frac{(\mathbf{C}\beta)'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}\mathbf{C}\beta}{2\sigma^2}$).
- Notice that if $\mathbf{C}\beta = \mathbf{0}$ is true, λ defined above = 0, giving F ~ F(q, n-rank(X)).

5 Appendix: Tangled R code

```
library(MASS); library(xtable)
     lvector <- function(x, dig = 2, dsply=rep("f",ncol(x)+1)) {
      x <- xtable(x, align=rep("",ncol(x)+1),display=dsply,digits=dig) # We repeat empty string 6 t
      print(x, floating=FALSE, tabular.environment="pmatrix",
        hline.after=NULL, include.rownames=FALSE, include.colnames=FALSE)
      }

   #Variables from Problem 2 of HW3:
     Y1 <- matrix(c(2, 1, 4, 6, 3, 5), nrow=6, ncol=1)
     X1 <- matrix(c(rep(1,6),
                    1,1,0,0,0,0,
                    0,0,1,0,0,0,
                    0,0,0,1,0,0,
                    0,0,0,0,1,1),nrow = 6,byrow=FALSE)

   #Variables from Problem 4 of HW3:
   data(Boston)
   YB = as.matrix(Boston$medv)
   XB = as.matrix(Boston[,c('crim','nox','rm','age','dis')])
   XB = cbind(rep(1,dim(Boston)[1]),XB)
   bhatB <- ginv(t(XB)%*%XB) %*% t(XB) %*% YB
   YhatB <- XB %*% bhatB
   errB <- YB - YhatB
   sigsqhatB <- t(errB) %*% errB / (dim(XB)[1] - qr(XB)$rank)

   #Another dataset tested:
   X511 <- matrix(c(rep(1,9),rep(c(rep(0,7),1),3),1,1),7,5)
   Y511 <- c(2,1,4,6,3,5,4)

   #functions for calculating estimates:
```

```
sigmacalc <- function(Y, X, alpha){
Yh <- X %*% ginv(t(X) %*% X) %*% t(X) %*% Y
SSE <- t(Y-Yh) %*% (Y-Yh)

lowerchi <- qchisq(alpha/2, df=(length(Y) - qr(X)$rank))
upperchi <- qchisq(1-alpha/2, df=(length(Y) - qr(X)$rank))

return(c(sqrt(SSE/upperchi),sqrt(SSE/lowerchi)))
}


cbetacalc <- function(Y,X, alpha, ct){
Yh <- X %*% ginv(t(X) %*% X) %*% t(X) %*% Y
SSE <- t(Y-Yh) %*% (Y-Yh)
MSE <- SSE/(length(Y) - qr(X)$rank)
quad <- t(ct)%*%ginv(t(X)%*%X)%*%ct

tstar <- qt(1-alpha/2, length(Y) - qr(X)$rank)
pm <- tstar * sqrt(quad) * sqrt (MSE)

ctbhat <- t(ct)%*%ginv(t(X)%*%X)%*%t(X)%*%Y

return(c(ctbhat-pm,ctbhat+pm))
}

#F-test function:

Cbetahatd <- function (Y,X, ct, d = 0){

CGCinv <- ginv(t(ct)%*%ginv(t(X)%*%(X))%*%ct)
CBhat <- t(ct)%*%ginv(t(X)%*%X)%*%t(X)%*%Y
Q <- t(CBhat - d)%*%CGCinv%*%(CBhat-d)
MSH <- Q/qr(ct)$rank

Yhat <- X %*% ginv(t(X)%*%X)%*%t(X)%*%Y
SSE <- t(Y-Yhat)%*%(Y-Yhat)
MSE <- SSE/(length(Y) - qr(X)$rank)

F <- MSH/MSE

return(1-pf(F, qr(ct)$rank, length(Y)-qr(X)$rank))

}

# Prediction limits

predict <- function (Y, X, alpha, ct, n=1){
Yh <- X %*% ginv(t(X) %*% X) %*% t(X) %*% Y
SSE <- t(Y-Yh) %*% (Y-Yh)
MSE <- SSE/(length(Y) - qr(X)$rank)
```

```
    quad <- t(ct)%*%ginv(t(X)%*%X)%*%ct
    gamma <- 1/n
    tstar <- qt(1-alpha/2, length(Y) - qr(X)$rank)
    pm <- tstar * sqrt(gamma+quad) * sqrt (MSE)

    ctbhat <- t(ct)%*%ginv(t(X)%*%X)%*%t(X)%*%Y

    return(c(ctbhat-pm, ctbhat+pm))
}

  #1e
  predict(Y1, X1, .1, c(1,0,1,0,0), 10)

  #1f
  predict(Y1, X1, .1, (c(1,1,0,0,0) - c(1,0,1,0,0)), 2)

  #1g
  G <- t(matrix(c(0,1,-1,0,0,
                  0,1,0,-1,0,
                  0,1,0,0,-1),nrow=3,ncol=5, byrow=TRUE))
  Cbetahatd(Y1,X1,G,c(0,0,0))

  #1h
  H <- t(matrix(c(0, 1, -1, 0, 0, 0, 0, 1, -1, 0), nrow=2, ncol=5, byrow=T))

  Cbetahatd(Y1,X1,H,c(10,0))

  #2a
  sigmacalc(YB,XB, .1)

  #2b
  cbetacalc(YB,XB, .1, XB[1,])

  #2c
  cbetacalc(YB,XB, .1, (XB[1,]-XB[2,]))

  #2d
  Cbetahatd(YB,XB, (XB[1,]-XB[2,]))

  #2e
  predict(YB,XB, .1, c(0,0.005,0.45,7,45,6), 1)

  #2f
  Cbetahatd(YB,XB, c(0,0,1,0,1,0))

  #3

#Find SSR in the full model.
bhat_B <- ginv(t(XB)%*%XB)%*%t(XB)%*%YB
SSR_Bf <- t(bhat_B) %*% t(XB) %*% YB - (length(YB)*(mean(YB))^2)
```

```
#create reduced model design matric and X1_B and estimator bhat1_B
X1_B <- XB[,-c(3,5)]
bhat1_B <- ginv(t(X1_B)%*%X1_B) %*% t(X1_B) %*% YB
SSR_Br <- t(bhat1_B) %*% t(X1_B) %*% YB - (length(YB)*(mean(YB))^2)

YhatB <- XB%*%bhat_B
SSE_B <- t(YB -YhatB)%*%(YB-YhatB)

F_2f <- ((SSR_Bf - SSR_Br)/2)/(SSE_B/(length(YB) - qr(XB)$rank))

pf_2f <- pf(F_2f, 2, (length(YB)-(qr(XB)$rank)), lower.tail=FALSE)

fractions(ginv(t(C1g)%*%ginv(t(X1)%*%X1)%*%C1g))
```