620-371: Linear Models

Assignment 2

Due: Monday, 4th May, 2009

This assignment is worth 5% of your total mark.

You may use R (and only R) for this assignment, but for matrix calculations and statistical distributions only. You may not use the `lm` function. If you do use R, include a computer printout of your commands and R output.

An experiment is conducted to estimate the demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained:

Cost ($\$k$)	Unemployment rate (%)	Interest rate (%)	Cars sold ($\times 10^3$)
9.0	10.0	4.0	6.5
5.5	9.0	7.0	5.9
9.0	12.0	5.0	8.0
9.8	11.0	6.2	9.0
14.5	12.0	5.8	10.0
8.0	13.7	3.9	10.8

1. (a) [2 marks] Calculate the least squares estimator, $\mathbf{b}$.

   Solution:

```
> library(car)
> X <- matrix(c(rep(1, 6), 9, 5.5, 9, 9.8, 14.5, 8, 10, 9, 12,
+     11, 12, 13.7, 4, 7, 5, 6.2, 5.8, 3.9), 6, 4)
> y <- as.vector(c(6.5, 5.9, 8, 9, 10, 10.8))
> b <- inv(t(X) %*% X) %*% t(X) %*% y
> b
           [,1]
[1,] -7.6023304
[2,]  0.1286809
[3,]  1.1303420
[4,]  0.3796061
```

   (b) [2 marks] In this year, the unemployment rate is 7% and the interest rate is 8%. What price should a dealer put on a car to expect to sell 7500 cars?

   Solution: If c is the cost, we want

$$\begin{bmatrix} 1 & c & 7 & 8 \end{bmatrix} \mathbf{b} = 7.5.$$

```
> (7.5 - b[1] - 7 * b[3] - 8 * b[4])/b[2]
```

```
[1] 32.27432
```

Therefore the price should be $32,274.

(c) [3 marks] Find a 95% confidence interval for the average number of $10,000 cars sold in a year which has unemployment rate 4.5% and interest rate 7%.

Solution:

```
> s2 <- sum((y - X %*% b)^2)/(6 - 4)
> xst <- as.vector(c(1, 10, 4.5, 7))
> halfwidth <- qt(0.975, df = 6 - 4) * sqrt(s2) * sqrt(t(xst) %*%
+     inv(t(X) %*% X) %*% xst)
> c(xst %*% b - halfwidth, xst %*% b + halfwidth)
```

```
[1] -6.703628  9.560148
```

We would expect to sell between -6703 cars (!) and 9560 cars.

2. (a) [2 marks] Test for model adequacy.

Solution:

```
> SSReg <- t(y) %*% X %*% inv(t(X) %*% X) %*% t(X) %*% y
> SSRes <- s2 * (6 - 4)
> Fstat <- (SSReg/4)/(SSRes/(6 - 4))
> Fstat
```

```
         [,1]
[1,] 143.2817
```

```
> pf(Fstat, 4, 6 - 4, lower.tail = FALSE)
```

```
            [,1]
[1,] 0.006942893
```

The model is clearly adequate.

(b) [5 marks] We want to find the best model for this data. Starting with an empty model, perform tests at the 5% level to determine if each parameter should be added to the model, in the order intercept, cost, unemployment, interest rate. If a parameter should be added, add it for the remaining tests; otherwise, leave it out. Using this approach, what is the best model?

Solution:

```
> X0 <- X[, 1]
> R0 <- t(y) %*% X0 %*% inv(t(X0) %*% X0) %*% t(X0) %*% y
> R0
```

```
         [,1]
[1,] 420.0067
```

```
> Fstat <- (R0/1)/(SSRes/(6 - 4))
> Fstat
```

```
         [,1]
[1,] 550.6204
```

```
> pf(Fstat, 1, 6 - 4, lower.tail = FALSE)
```

```
            [,1]
[1,] 0.001811201
```

We should add the intercept parameter.

```
> X1 <- X[, c(1, 2)]
> R1 <- t(y) %*% X1 %*% inv(t(X1) %*% X1) %*% t(X1) %*% y
> R10 <- R1 - R0
> R10
```

```
         [,1]
[1,] 5.646241
```

```
> Fstat <- (R10/1)/(SSRes/(6 - 4))
> Fstat
```

```
        [,1]
[1,] 7.40211
```

```
> pf(Fstat, 1, 6 - 4, lower.tail = FALSE)
```

```
         [,1]
[1,] 0.112711
```

We should not add the cost parameter.

```
> X2 <- X[, c(1, 3)]
> R2 <- t(y) %*% X2 %*% inv(t(X2) %*% X2) %*% t(X2) %*% y
> R20 <- R2 - R0
> R20
```

```
         [,1]
[1,] 15.51469
```

```
> Fstat <- (R20/1)/(SSRes/(6 - 4))
> Fstat
```

```
         [,1]
[1,] 20.33945
```

```
> pf(Fstat, 1, 6 - 4, lower.tail = FALSE)
```

```
           [,1]
[1,] 0.04581329
```

We should add an unemployment parameter.

```
> X3 <- X[, c(1, 3, 4)]
> R3 <- t(y) %*% X3 %*% inv(t(X3) %*% X3) %*% t(X3) %*% y
> R32 <- R3 - R2
> R32
```

```
         [,1]
[1,] 1.057768
> Fstat <- (R32/1)/(SSRes/(6 - 4))
> Fstat
         [,1]
[1,] 1.386713
> pf(Fstat, 1, 6 - 4, lower.tail = FALSE)
          [,1]
[1,] 0.3601118
```

We should not add the interest rate parameter. Therefore the best model uses only the intercept term and the unemployment rate.

(c) [3 marks] Simultaneously test the hypotheses that the unemployment rate is not relevant to the cars sold and that the parameters corresponding to cost and interest rate are equal.

Solution: We are testing the hypothesis $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

C is of rank 2, so we can apply the test for the general linear hypothesis.

```
> C <- matrix(c(0, 0, 0, 1, 1, 0, 0, -1), 2, 4)
> Fstat <- (t(C %*% b) %*% inv(C %*% inv(t(X) %*% X) %*% t(C)) %*%
+     C %*% b/2)/(SSRes/(6 - 4))
> Fstat
         [,1]
[1,] 9.325361
> pf(Fstat, 2, 6 - 4, lower.tail = FALSE)
          [,1]
[1,] 0.0968489
```

We cannot reject these hypotheses.

3. [5 marks] We can derive a joint confidence region for the parameters of a full rank linear model,

$$(\mathbf{b} - \boldsymbol{\beta})^T X^T X (\mathbf{b} - \boldsymbol{\beta}) \leq ps^2 f_\alpha.$$

Use this joint confidence region to derive a test for the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. Show that this test is equivalent to the existing test for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$.

Solution: We accept H_0 if and only if $\boldsymbol{\beta}^*$ lies in the joint confidence region. This means that we accept it if and only if

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X (\mathbf{b} - \boldsymbol{\beta}^*)^T \leq ps^2 f_\alpha.$$

But the existing test accepts H_0 if and only if

$$\begin{aligned} \frac{(\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*)/p}{SS_{Res}/(n-p)} &\leq f_\alpha \\ (\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) &\leq p\frac{SS_{Res}}{n-p} f_\alpha \\ (\mathbf{b}-\boldsymbol{\beta}^*)^T X^T X(\mathbf{b}-\boldsymbol{\beta}^*) &\leq ps^2 f_\alpha. \end{aligned}$$

Therefore the two tests are equivalent.