

MAST30025: Linear Statistical Models

Solutions to Week 10 Lab

1. Consider again the data from Question 5 from the Week 8 lab. In a manufacturing plant, filters are used to remove pollutants. We are interested in comparing the lifespan of 5 different types of filters. Six filters of each type are tested, and the time to failure in hours is given in the dataset `filters` (on the website, in `csv` format).

- (a) Calculate s^2 .

Solution:

```
> n <- length(y)
> library(Matrix)
> r <- rankMatrix(X)[1]
> e <- y - X %>% b
> (s2 <- sum(e^2)/(n-r))

[1] 15304.2
```

- (b) Calculate a 95% confidence interval for the difference in lifespan between filter types 3 and 4.

Solution:

```
> tt <- c(0,0,0,1,-1,0)
> hw <- qt(0.975,df=n-r) * sqrt(s2 * t(tt) %>% XtXc %>% tt)
> c(tt %>% b - hw, tt %>% b + hw)

[1] -338.43399 -44.23268
```

- (c) Show that the hypothesis that the filters all have the same lifespan is testable.

Solution:

```
> C <- matrix(c(0,1,-1,0,0,0,0,0,1,-1,0,0,0,0,0,1,-1,0,0,0,0,0,1,-1),4,6,byrow=TRUE)
> C %>% XtXc %>% t(X) %>% X

      [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 5.551115e-17 1 -1 0 0 0
[2,] 5.551115e-17 0 1 -1 0 0
[3,] 5.551115e-17 0 0 1 -1 0
[4,] 5.551115e-17 0 0 0 1 -1
```

- (d) Test this hypothesis, using matrix theory.

Solution:

```
> (Fstat <- (t(C%b) %>% solve(C %>% XtXc %>% t(C)) %>% C %>% b/4)/s2)

      [,1]
[1,] 3.318776
> pf(Fstat, 4, n-r, lower.tail=FALSE)

      [,1]
[1,] 0.02599945
```

We reject the hypothesis at a 5% level.

- (e) Test the same hypothesis using the `linearHypothesis` function from the `car` package.

Solution:

```
> filters$type <- factor(filters$type)
> model <- lm(life ~ type, data = filters)
> library(car)
> C2 <- matrix(c(0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1),4,5,byrow=TRUE)
> linearHypothesis(model, C2, rep(0,4))
```

Linear hypothesis test

Hypothesis:

type2 = 0

type3 = 0

type4 = 0

type5 = 0

Model 1: restricted model

Model 2: life ~ type

```

      Res.Df    RSS Df Sum of Sq      F Pr(>F)
1         29 585770
2         25 382605   4    203165 3.3188 0.026 *
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

2. Derive a formula for a prediction interval for $\mathbf{t}^T \boldsymbol{\beta}$ in the less than full rank model.

Solution: $\mathbf{t}^T \mathbf{b}$ has a normal distribution with mean $\mathbf{t}^T \boldsymbol{\beta}$ and variance $\sigma^2 \mathbf{t}^T (X^T X)^c \mathbf{t}$. The response for a particular sample with design variables \mathbf{t} is $\mathbf{y}^* = \mathbf{t}^T \boldsymbol{\beta} + \varepsilon$, where ε has mean 0 and variance σ^2 . Therefore $\mathbf{t}^T \mathbf{b} - \mathbf{y}^*$ has a normal distribution with mean 0 and variance $\sigma^2 (1 + \mathbf{t}^T (X^T X)^c \mathbf{t})$. Dividing by s/σ to get a t distribution and re-arranging gives the prediction interval

$$\mathbf{t}^T \mathbf{b} \pm t_{\alpha/2} s \sqrt{1 + \mathbf{t}^T (X^T X)^c \mathbf{t}},$$

using a t distribution with $n - r$ degrees of freedom.

3. Consider a one-way classification model with three levels. To test $\tau_1 = \tau_2 = \tau_3$ we could use either of the following matrices:

$$C_1 = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Show that the test statistics formed using these two matrices are the same.

Hint: find A such that $AC_1 = C_2$.

Solution: Let $A = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$, then $AC_1 = C_2$. The numerator for the test statistic using C_2 is

$$\begin{aligned} (C_2 \mathbf{b})^T [C_2 (X^T X)^c C_2^T]^{-1} C_2 \mathbf{b} / r(C_2) &= (AC_1 \mathbf{b})^T [AC_1 (X^T X)^c C_1^T A^T]^{-1} AC_1 \mathbf{b} / r(AC_1) \\ &= (C_1 \mathbf{b})^T A^T (A^T)^{-1} [C_1 (X^T X)^c C_1^T]^{-1} A^{-1} AC_1 \mathbf{b} / r(C_1) \\ &= (C_1 \mathbf{b})^T [C_1 (X^T X)^c C_1^T]^{-1} C_1 \mathbf{b} / r(C_1), \end{aligned}$$

which is the numerator for the test using C_1 . The denominator $s^2/(n - r)$ is the same in each case.

4. An industrial psychologist is investigating absenteeism among production-line workers, based on different types of work hours: (1) 4-day week with a 10-hour day, (2) 5-day week with a flexible 8-hour day, and (3) 5-day week with a structured 8-hour day. A study is conducted and the following data obtained of the average number of days missed:

	Work plan		
	1	2	3
Mean	9	6.2	10.1
Number	100	85	90

They also find $s^2 = 110.15$.

- (a) Test the hypothesis that the work plan has no effect on the absenteeism.

Solution:

```

> b <- c(0,9,6.2,10.1)
> s2 <- 110.15
> n <- c(100,85,90)
> r <- 3
> XtXc <- diag(c(0,1/n))
> (C <- matrix(c(0,1,-1,0,0,0,1,-1),2,4,byrow=T))

      [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1
> (Fstat <- t(C%*%b)%*%solve(C%*%XtXc%*%t(C))%*%C%*%b/2/s2)

      [,1]
[1,] 3.200371
> pf(Fstat,2,sum(n)-r,lower=F)

      [,1]
[1,] 0.04228613

```

Therefore we reject the null hypothesis at a 5% level: work plan has an effect on absenteeism.

- (b) Test the hypothesis that work plans 1 and 3 have the same rate of absenteeism.

Solution:

```

> (C <- matrix(c(0,1,0,-1),1,4,byrow=T))

      [,1] [,2] [,3] [,4]
[1,]    0    1    0   -1
> (Fstat <- t(C%*%b)%*%solve(C%*%XtXc%*%t(C))%*%C%*%b/1/s2)

      [,1]
[1,] 0.5203431
> pf(Fstat,1,sum(n)-r,lower=F)

      [,1]
[1,] 0.471315

```

We cannot reject the null hypothesis.