

620-371: Linear Models

Assignment 3

Due: Monday, 25th May, 2009

This assignment is worth 5% of your total mark.

You may use R (and only R) for this assignment, but only for questions 1 and 2e. You may not use the `lm` function for question 1, but you may use it for question 2e. If you do use R, include a computer printout of your commands and R output.

1. We are interested in examining the yield of tomato plants that have been grown with certain types of fertiliser. A study is conducted and the following data obtained:

Fertiliser		
1	2	3
43	33	54
45	37	54
47	38	57
46	35	
48		

We fit the model

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

where μ is the overall mean and τ_i is the effect of using the i th fertiliser.

- (a) [2 marks] Find a conditional inverse for $X^T X$, using the algorithm given in the lecture slides.

Solution:

```
> X <- matrix(c(rep(1, 17), rep(0, 12), rep(1, 4), rep(0, 12),
+ rep(1, 3)), 12, 4)
> y <- as.vector(c(43, 45, 47, 46, 48, 33, 37, 38, 35, 54, 54,
+ 57))
> t(X) %*% X
```

```
      [,1] [,2] [,3] [,4]
[1,]   12    5    4    3
[2,]    5    5    0    0
[3,]    4    0    4    0
[4,]    3    0    0    3
```

A conditional inverse is

$$(X^T X)^c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

- (b) [3 marks] Find two solutions to the normal equations, using only the conditional inverse you found in question 1a.

Solution:

```
> XtXc <- diag(c(0, 1/5, 1/4, 1/3))
> I <- diag(rep(1, 4))
> b <- XtXc %*% t(X) %*% y
> b
```

```
      [,1]
[1,]  0.00
[2,] 45.80
[3,] 35.75
[4,] 55.00
```

```
> b2 <- b + (I - XtXc %*% t(X) %*% X) %*% as.vector(c(1, 0, 0,
+      0))
> b2
```

```
      [,1]
[1,]  1.00
[2,] 44.80
[3,] 34.75
[4,] 54.00
```

- (c) [2 marks] Is $\mu + \tau_1 - \tau_2 + \tau_3$ estimable?

Solution: Yes; $\mu + \tau_1$ is estimable as it is an element of $X\beta$, and $\tau_2 - \tau_3$ is a treatment contrast, so their sum is estimable.

- (d) [3 marks] Find a 95% confidence interval for the estimable quantity $\tau_2 - \tau_3$.

Solution:

```
> n <- 12
> r <- 3
> tt <- as.vector(c(0, 0, 1, -1))
> s2 <- sum((y - X %*% b)^2)/(n - r)
> hw <- qt(0.975, df = n - r) * sqrt(s2 * t(tt) %*% XtXc %*% tt)
> c(tt %*% b - hw, tt %*% b + hw)
```

```
[1] -22.68384 -15.81616
```

- (e) [2 marks] Test the hypothesis that fertiliser has no effect on yield.

Solution: This hypothesis can be written as $H_0 : C\beta = \mathbf{0}$, where

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

```
> library(car)
> C <- matrix(c(0, 0, 1, 0, -1, 1, 0, -1), 2, 4)
> Fstat <- (t(C %*% b) %*% inv(C %*% XtXc %*% t(C)) %*% C %*% b/2)/s2
> Fstat
```

```
      [,1]
[1,] 81.6076
> pf(Fstat, 2, n - r, lower.tail = FALSE)
```

```
      [,1]
[1,] 1.705178e-06
```

We reject the null hypothesis firmly.

2. We study the amount of rotting of a potato exposed to a variety of levels of oxygen, and a variety of temperatures. A small experiment is conducted and the following data obtained:

Temperature	Oxygen level		
	1	2	3
10	13	10	15
	11	4	2
	3	7	7
16	26	15	20
	19	22	24
	24	18	8

We fit the model

$$y_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}, \quad (1)$$

where μ is the overall mean, τ_i is the effect of the i th oxygen level, and β_j is the effect of the j th temperature level. A model was fitted in R using the `lm` function, from which the following output was derived:

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> model <- lm(rot ~ oxygen.f + temp.f, data = potato)
> summary(model)
```

Call:

```
lm(formula = rot ~ oxygen.f + temp.f, data = potato)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-10.4444  -2.8611   0.4444   3.0278   8.1111
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.222	2.410	4.242	0.000820 ***
oxygen.f2	-3.333	2.951	-1.130	0.277660
oxygen.f3	-3.333	2.951	-1.130	0.277660
temp.f16	11.556	2.410	4.796	0.000285 ***

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

Residual standard error: 5.111 on 14 degrees of freedom
Multiple R-squared: 0.6382, Adjusted R-squared: 0.5607
F-statistic: 8.233 on 3 and 14 DF, p-value: 0.002105

> anova(model)

Analysis of Variance Table

Response: rot

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
oxygen.f	2	44.44	22.22	0.8505	0.4481124
temp.f	1	600.89	600.89	22.9988	0.0002849 ***
Residuals	14	365.78	26.13		

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

> linear.hypothesis(model, c(0, 1, 0, -1), 0)

Linear hypothesis test

Hypothesis:

oxygen.f2 - temp.f16 = 0

Model 1: rot ~ oxygen.f + temp.f

Model 2: restricted model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	14	365.78				
2	15	764.80	-1	-399.02	15.272	0.001577 **

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

- (a) [2 marks] What information do you gather about the parameters of the model in equation 1, from the coefficients of the fitted R model?

Solution: From the coefficients, we deduce that

$$\begin{aligned}\mu + \tau_1 + \beta_1 &\approx 10.22 \\ \tau_2 - \tau_1 &\approx -3.33 \\ \tau_3 - \tau_1 &\approx -3.33 \\ \beta_2 - \beta_1 &\approx 11.56\end{aligned}$$

- (b) [1 mark] Should we accept or reject the hypothesis that oxygen level has no effect on rot?

Solution: We should accept it; the p -value is 0.448.

- (c) [1 mark] Should we accept or reject the hypothesis that temperature has no effect on rot?

Solution: We should reject it; the p -value is 0.00028.

- (d) [2 marks] What hypothesis is the **linear.hypothesis** function testing, given in terms of the parameters of the model in equation 1?

Solution: The hypothesis is testing $\tau_2 - \tau_1 = \beta_2 - \beta_1$; in other words, whether the difference in effect between the temperature levels is equal to the difference in effect between the first two oxygen levels.

- (e) [2 marks] Using R, test for the presence of interaction between the factors.

Solution:

```
> model2 <- lm(rot ~ oxygen.f * temp.f, data = potato)
> anova(model2)
```

Analysis of Variance Table

Response: rot

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
oxygen.f	2	44.44	22.22	0.7634	0.487453
temp.f	1	600.89	600.89	20.6412	0.000674 ***
oxygen.f:temp.f	2	16.44	8.22	0.2824	0.758816
Residuals	12	349.33	29.11		

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

There clearly is no interaction — the p -value is 0.76.

3. [4 marks] Show that for any matrix A , the matrix $A(A^T A)^c A^T$ is invariant to the choice of a conditional inverse (i.e. unique). You may use the property of a general matrix M that if $M^T M = 0$, then $M = 0$. (*Hint: Let $(A^T A)_1^c$ and $(A^T A)_2^c$ be two different conditional inverses for $A^T A$ and show that $A(A^T A)_1^c A^T - A(A^T A)_2^c A^T = 0$.)*

Solution: Let $M = A(A^T A)_1^c A^T - A(A^T A)_2^c A^T$. Then

$$\begin{aligned}
M^T M &= (A(A^T A)_1^c A^T - A(A^T A)_2^c A^T)^T (A(A^T A)_1^c A^T - A(A^T A)_2^c A^T) \\
&= A(A^T A)_1^c A^T A(A^T A)_1^c A^T - A(A^T A)_2^c A^T A(A^T A)_1^c A^T \\
&\quad - A(A^T A)_1^c A^T A(A^T A)_2^c A^T + A(A^T A)_2^c A^T A(A^T A)_2^c A^T \\
&= A(A^T A)_1^c A^T - A(A^T A)_1^c A^T - A(A^T A)_2^c A^T + A(A^T A)_2^c A^T \\
&= 0.
\end{aligned}$$

Hence $M = 0$ and the result follows.