

**The University of Melbourne  
Assessment, Semester 1, 2007**

**Department of Mathematics and Statistics  
620–371 Linear Models**

**Exam duration: Three hours  
Reading time: fifteen minutes  
This paper has nine (9) pages, including this one.**

**Authorised materials:**

Hand-held electronic calculators may be used.

**Instructions to invigilators:**

Statistical tables will be supplied.

**Instructions to students:**

There are seven (7) questions. All questions may be attempted. They carry weights as shown in the brackets after the question statement. These marks total 115.

*This paper is to be lodged with the Baillieu Library*

1. (a) For the general linear model  $y = \mathbf{X}\beta + \varepsilon$  where  $\varepsilon \stackrel{d}{\sim} N(0, \sigma^2 \mathbf{I})$ , the least squares estimate of  $\beta$  is given as the solution(s) of the “normal equations”,

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'y$$

For a given situation with

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}$$

it is found that

$$\hat{\beta}^{(1)} = \begin{bmatrix} 31 \\ 15 \\ 25 \\ 18 \\ 19 \\ 16 \end{bmatrix} \quad \text{and} \quad \hat{\beta}^{(2)} = \begin{bmatrix} 20 \\ 4 \\ 36 \\ 7 \\ 30 \\ 5 \end{bmatrix}$$

are two solutions of the normal equations.

- i. Are there likely to be more solutions of the normal equations and, if so, how many?
  - ii. The rank of the design matrix,  $\mathbf{X}$ , is 5, 6 or 7. State, with reasons, which it must be.
  - iii. Exactly one of the following three parametric functions is estimable. Which is it, and why?
    - A.  $\beta_3$
    - B.  $\beta_3 + \beta_4$
    - C.  $\beta_3 - \beta_4$
  - iv. Parametric functions such as those given in (iii.) can be used either to impose constraints on the normal equations (to enable a particular solution to be found) or to form (part of) a null hypothesis (to be tested). State, with reasons, which of the three parametric functions in (iii.) could be used:
    - A. to impose a constraint on the normal equations;
    - B. (as part of) a null hypothesis.
- (b) For the simple straight line regression model  $y_i = \alpha + \beta x_i + e_i$ ,  $i = 1, \dots, n$ , explain how you could test the hypothesis  $\alpha = \beta$  using:
- i. a  $t$ -test;
  - ii. an  $F$ -test (using the general form of the  $F$ -test for linear models).

Give details of any models that you would fit, and how you would obtain all of the information required to complete the tests from  $\mathbf{R}$ .

[9 + 5 = 14 marks]

2. The following data were obtained from a study of the effect of different substances (such as water, milk, etc) on the survival of *Salmonella typhimurium*. The response variable ( $y$ ) was taken to be the logarithm of the number of live cells per ml, after 24 hours.

A randomised block design was used (each block consisted of a different batch of *Salmonella typhimurium* cells) with each of the five materials used three times in each block. However, due to difficulties with the experiment, seven observations were not recorded, leaving a total of 38 observations for the analysis.

Substance	batch (block)		
	1	2	3
1	8.1	7.9	7.0
	9.1	8.0	7.6
	7.8	8.1	
2	7.2	6.9	5.7
	5.7	6.4	6.8
	6.6		7.3
3	6.6	4.5	6.0
	6.4	5.3	5.1
		5.3	4.1
4	5.2	4.7	4.6
		5.0	4.6
		3.8	
5	5.2	3.2	4.1
	4.4	3.9	4.1
	4.9		4.2

Model [ $\text{lm}(y \sim \dots)$ ]	Deviance
1	84.35
batch.f	77.77
substance.f	14.24
batch.f+substance.f	10.57
batch.f*substance.f	7.34

[**Note:** batch.f is a factor with three levels, while substance.f is a factor with five levels.]

- Explain what was probably done, and observed, that led to the log transformation being used.
- Carry out a test to show that the additive model provides an adequate fit to the data.
- Complete the analysis of the data, including tests to determine the significance of the factors and the use of Tukey's multiple comparisons to determine which levels of water activity differ significantly (at the 5% level), and state your conclusions.

[For the Tukey comparisons, just do enough to decide which differences are statistically significant, and to convince me that you know what you are doing.]

```

> summary(lm(y~substance.f+batch.f))$coef
              Estimate Std. Error   t value    Pr(>|t|)
(Intercept)  8.3747300  0.2442293  34.290436 3.225755e-26
substance.f2 -1.3761429  0.2934499  -4.689532 5.219554e-05
substance.f3 -2.4532397  0.2935022  -8.358505 1.927466e-09
substance.f4 -3.1584233  0.3183989  -9.919706 3.883874e-11
substance.f5 -3.7011429  0.2934499 -12.612520 9.585929e-14
batch.f2     -0.6832253  0.2386901  -2.862395 7.471072e-03
batch.f3     -0.6740821  0.2357323  -2.859524 7.524562e-03

> disp.s(lm(y~substance.f+batch.f))
[1] "Matrix of estimated standard errors of differences"
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,] 0.0000000 0.4816465 0.4741635 0.4847276 0.4816465 0.4172350 0.4063300
[2,] 0.4816465 0.0000000 0.2935396 0.3202049 0.2920193 0.3690661 0.3849776
[3,] 0.4741635 0.2935396 0.0000000 0.3164925 0.2935396 0.3879223 0.3944688
[4,] 0.4847276 0.3202049 0.3164925 0.0000000 0.3202049 0.4213141 0.4166250
[5,] 0.4816465 0.2920193 0.2935396 0.3202049 0.0000000 0.3690661 0.3849776
[6,] 0.4172350 0.3690661 0.3879223 0.4213141 0.3690661 0.0000000 0.2315278
[7,] 0.4063300 0.3849776 0.3944688 0.4166250 0.3849776 0.2315278 0.0000000

```

- (d) It is claimed that survival depends solely upon the so-called “water activity” (the relative availability of water) of a substance. Water activity is a numerical quantity, typically taking values between 0.5 and 1, but the values were not reported in this study. Explain why it would be desirable to use these values in the analysis, and explain how you would carry out the analysis — what models would you fit, what hypotheses would you test and under what circumstances would you conclude that the claim, namely that survival depends (only) on the level of water activity, might be reasonable?

[2 + 2 + 6 + 4 = 14 marks]

3. The data for this question come from a study conducted in forests near Ballarat. The main purpose of the study was to investigate the effects of three tree thinning “treatments”, which are thought to improve tree growth by reducing competition, with other trees. The response variable considered here is the diameter of the trees at breast height (dbh), in addition to which the distance of each tree to its nearest neighbour was also recorded. Data were available for a total of 48 trees; 20 from treatment 1, 14 from treatment 2 and 14 from treatment 3.

The following deviances were obtained using **R**, where *treat.f* denotes the thinning treatments (treated as a factor) and *d* denotes the distance of a tree to its nearest neighbour.

Model	Deviance
1 1	840524
2 treat.f	490097
3 <i>d</i>	720912
4 treat.f + <i>d</i>	457923
5 treat.f* <i>d</i>	378548

- (a) i. Give the parametric form (eg  $y_{ij} = \alpha_i + \beta d_{ij} + e_{ij}$ ), and the degrees of freedom for the deviance, for each of the five models given in the table.

- ii. For models 2 to 5 (only), which of the deviances would be **unchanged** if the term “1” were to be subtracted from the **R** specification (eg model 2 becomes  $\text{lm}(y \sim \text{treat.f} - 1)$ ).
  - iii. Which pairs of the five models, if any, **cannot** be formally compared using an  $F$ -test?
  - iv. Use  $F$ -tests to determine which is the most appropriate model for these data.
- (b) The following output was obtained using a model that is equivalent to one of the models in the above table, but with a different **R** model specification.

```
> summary(klm.1)
```

```
Call: lm(formula = dbh ~ treat.f + treat.f:d - 1)
```

Coefficients:

	Value	Std. Error	t value	Pr(> t )
treat.f1	544.9737	103.7595	5.2523	0.0000
treat.f2	809.7610	72.4879	11.1710	0.0000
treat.f3	347.9560	98.6306	3.5279	0.0010
treat.f1d	1.4136	0.6029	2.3446	0.0239
treat.f2d	-0.4214	0.4967	-0.8485	0.4009
treat.f3d	1.8226	0.7344	2.4818	0.0172

```
> disp.s(klm.1)
```

```
[1] "Matrix of estimated standard errors of differences"
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,]  0.0000 126.5722 143.1573 104.3497 103.7606 103.7620
[2,] 126.5722  0.0000 122.4030  72.4904  72.9533  72.4916
[3,] 143.1573 122.4030  0.0000  98.6324  98.6318  99.3404
[4,] 104.3497  72.4904  98.6324  0.0000  0.7811  0.9501
[5,] 103.7607  72.9533  98.6318  0.7811  0.0000  0.8865
[6,] 103.7621  72.4916  99.3404  0.9501  0.8865  0.0000
```

- i. Which of models 1 – 5 is the model that was used equivalent to?
- ii. Which of the following would not be the same for the two **R** models:  $y$ ,  $\hat{y}$ ,  $\hat{\beta}$ , the deviance, the residuals, the standardised residuals?
- iii. Based solely on this output, it was suspected that there was no difference (at all) between treatments 1 and 3. Carry out appropriate tests (two are required) to show how this suspicion may have arisen.
- iv. Give details of the model that you would need to fit in order to carry out a formal test of the suspected equivalence of treatments 1 and 3. The appropriate model was fitted and resulted in a deviance of 500431. Complete the test and state your conclusion.

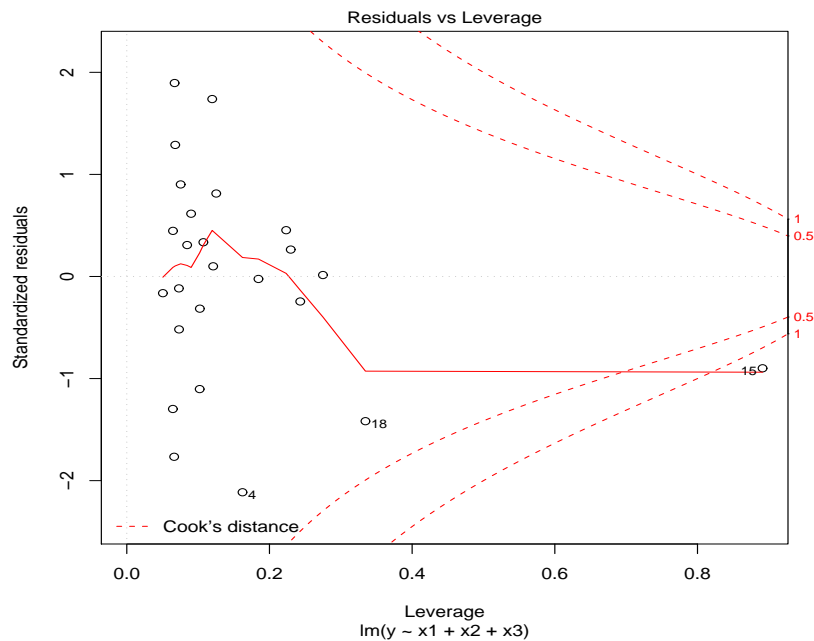
[12 + 10 = 22 marks]

4. A multiple regression model was fitted to a sample of 25 observations and the following **R** output obtained.

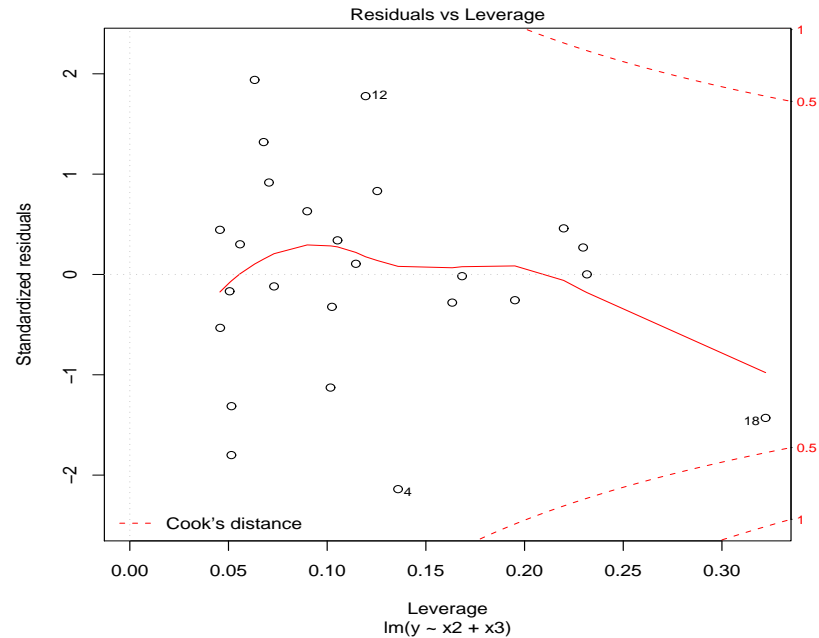
Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	13.0139	13.0139	340.20	1.881e-14 ***
x2	1	19.3655	19.3655	506.24	3.515e-16 ***
x3	1	8.7993	8.7993	230.02	8.694e-13 ***
Residuals	21	0.8033	0.0383		



- Explain, briefly, the role of each of the components in the above diagnostic plot (Leverage, Standardised residuals and Cook's distance) and comment on what the plot tells you about observation 15.
- On considering output not shown here, it was decided to omit variable  $x_1$  from the model. What additional output was looked at, what do you think it showed, and why was removing  $x_1$  not inconsistent with the highly significant  $P$ -value for  $x_1$  in the above output.
- From the diagnostics given for the regression of  $y$  on  $x_2$  and  $x_3$ , below, what can you now conclude about observation 15.



- (d) The following **R** output was obtained for the (multiple) regression of  $y$  on  $x_2$  and  $x_3$ . Find an estimate of, and a 95% prediction interval for, the value of  $y$  when  $x_2 = -2$  and  $x_3 = 0.5$ .

```
> summary(lm(y~x2+x3))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.97991	0.03865	77.10	< 2e-16 ***
x2	0.92921	0.05433	17.10	3.41e-14 ***
x3	1.82174	0.10146	17.96	1.25e-14 ***

Residual standard error: 0.1911 on 22 degrees of freedom  
 Multiple R-Squared: 0.9809, Adjusted R-squared: 0.9791  
 F-statistic: 563.8 on 2 and 22 DF, p-value: < 2.2e-16

```
> disp.v(lm(y~x2+x3))
```

[1] "Estimated covariance matrix of parameter vector"

	(Intercept)	x2	x3
(Intercept)	0.00149	0.00025	0.00010
x2	0.00025	0.00295	-0.00251
x3	0.00010	-0.00251	0.01029

[6 + 3 + 2 + 6 = 17 marks]

5. This question refers to a study into the effects of (four) denture cleansers on the hardness of (four) types of denture liner.

The purpose of a denture liner is to “achieve a more equal force distribution, to reduce localized pressure and to improve denture retention by engaging undercuts” – in brief, to improve denture comfort. To this end, the softer the liner the better it is likely to be.

The hardness of each of the (16) cleanser by liner combinations was measured 10 times at 24 hours and 10 times at six months, giving a total of 320 observations; 320 different samples were used, with one observation per sample.

- (a)
- i. How would you display the data?
  - ii. Complete the ANOVA table given below and state your conclusions (in terms of which effects are statistically significant).
  - iii. Explain, in detail, **why** and **how** you would use Tukey’s multiple comparisons here. In particular, give the relevant values of  $LSD_Q$ .

```
> liner.1 <- lm(Hardness~Cleanser*Liner*Time,data=liner.dat)
> anova(liner.1)
```

Response: Hardness

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cleanser		362			
Liner		63802			
Time		6229			
Cleanser:Liner		12			
Cleanser:Time		4			
Liner:Time		1587			
Cleanser:Liner:Time		6			
Residuals		908			

Just give a range (eg  $> 0.05$ ) for  $\text{Pr}(>F)$ .

- (b) Suppose now that, instead of 320 samples being used only once, there were 160 samples each measured twice; once at 24 hours and once at six months. Given that the “(between) samples  $SS = 64970$ ”, re-do the ANOVA table and state your conclusions in terms of which effects are (statistically) significant.

[13 + 5 = 18 marks]

6. A study was carried out to evaluate the effect of two drugs ( $A$  and  $B$ ) on lowering blood cholesterol levels. A total of 21 subjects were assigned, at random, to one of three groups so that each group consisted of seven subjects. Subjects in group I were given drug  $A$ , subjects in group II were given drug  $B$ , while subjects in group III were used as controls and given a placebo (ie tablets with no active ingredients). Two measurements were taken on each subject, one at the start of the study and one after the subjects had been on the drug, or placebo, for two months.

- (a)
- i. For each of the factors ‘drugs’, ‘subjects’ and ‘time’ (of measurement) state, with reasons, whether it should be treated as a fixed or a random effect.
  - ii. For each of the factors state whether it is nested or crossed with the other factors.

- iii. Describe how you would analyse the data from the experiment. Give the split-up of the degrees of freedom and state which mean-squares you would use to test the significance of the various effects and interactions.
  - iv. If the two drugs **do** affect cholesterol levels while the placebo **does not**, state, with reasons, which tests you might expect to be significant.
- (b) An alternative way to analyse the data from this study would be to use the initial cholesterol level for each subject as a covariate, with either
- the cholesterol level at the end of the study period or
  - the change in cholesterol level (final – initial level)

as the response variable. Explain why the two approaches (with the different response variables) are (essentially) equivalent. It may help to write down the (parametric) form of the model that would be fitted in the two approaches.

[11 + 3 = 14 marks]

7. (a) Explain the role of confounding in  $2^n$  factorial experiments; when is it used, and why?
- (b) A  $2^5$  factorial experiment (with factors  $A$ ,  $B$ ,  $C$ ,  $D$  and  $E$ ) is to be carried out in **blocks of eight plots**.
- i. If four blocks are available (giving a total of 32 observations), list effects that could be confounded with blocks, and give an allocation of treatments to blocks, so that no main effect or two-factor interaction is confounded. Describe two (substantially different) approaches that could be used to analyse the data, and give one point in favour of each of the two approaches. Explain how you would analyse the resultant data if you would like to test for as many effects (including interactions) as possible.
  - ii. Suppose now that an **additional** four blocks (of eight plots) are available for the experiment. Suggest effects that could be confounded in these additional blocks so that no main effect or two-factor interaction is confounded, and no effect is completely confounded. Explain how you would modify the analysis in (i) to account for the additional data.

[2 + 12 = 14 marks]