UNIVERSITY OF SURREY©

B. Sc. Undergraduate Programmes in Mathematical Studies
M. Math. Undergraduate Programmes in Mathematical Studies

Level HE2 Examination

Module MS236 GENERAL LINEAR MODELS

Time allowed – 2 hrs Autumn Semester 2007

Attempt THREE questions.
If any candidate attempts more than THREE questions only the best THREE solutions will be taken into account. Cambridge Tables will be provided.

SEE NEXT PAGE

Question 1

Experience with a certain type of plastic indicates that a relationship exists between the hardness (Y) (measured in Brinell units) of items moulded from the plastic and the elapsed time (x) in hours since termination of the moulding process. Twelve batches of the plastic were made and from each batch one test item was moulded and the hardness measured at some specific point in time. The results are shown below

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	32	48	72	64	48	16	40	48	48	24	80	56
y_i	230	262	323	298	255	199	248	279	267	214	359	305

(a) Assuming a simple linear regression model $Y_i = \alpha + \beta x_i + \varepsilon_i$, find the least squares estimates of α and β. [0]

(b) Show that the variances of the parameters $\hat{\beta}$ and $\hat{\alpha}$ are:

$$\text{Var}\left[\hat{\beta}\right] = \frac{\sigma^2}{S_{xx}} \text{ and } \text{Var}\left[\hat{\alpha}\right] = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Evaluate the standard error of the estimates for this data set. [0]

(c) Find a 95% confidence interval for the mean value of Y when $x = 56$. [5]
[Hint: The variance for the least squares predictor when $x = x_0$ is $\sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]$] [0]

(d) On analysing these data using R a researcher tried fitting successively the null model ($\text{lm}(y \sim 1)$)), a simple linear regression model ($\text{lm}(y \sim x)$) and a model including a quadratic term ($\text{lm}(y \sim x + x * x)$). The resulting deviances were 23379, 952.25 and 821.46. Write down an appropriate ANOVA table and test the hypothesis that the quadratic term is unnecessary.

Question 2

The multiple regression model with an intercept term and $p-1$ explanatory variables can be written in the form

$$Y = X\beta + \varepsilon$$

where X is an $n \times p$ matrix of rank p. Identify X and β for this model. [3]
Write down the normal equations and the least squares estimates in terms of X, Y and $\hat{\beta}$. Why is it necessary to assume that X is of rank p? [5]

A random sample of 15 applicants for a clerical job were given two aptitude tests with resultant scores X_1 and X_2. After a period of training each was given a Job Proficiency Score (Y). It is hoped to predict Proficiency Scores on the basis of the aptitude tests using a regression model. The table of deviances is as follows:

Variables included	**Deviance**
none	2562.1
X_1	110.3
X_2	196.4
X_1, X_2	15.5

SEE NEXT PAGE

Show that the overall regression on X_1 and X_2 is significant and that neither X_1 nor X_2 can be dropped. [7]

The statistician analyzing the data suggests that a linear model might not be appropriate since two moderate scores may indicate a different Proficiency than one high and one low score. He therefore includes the variable $X_3 = X_1X_2$.

Variables included	**Deviance**
X_3	16.1
X_1, X_3	15.8
X_2, X_3	16.0

Show that the regression on X_3 is significant but adding X_1 or X_2 does not significantly decrease the deviance using the initial estimate of σ^2 above. [5]

What criteria can be used to compare the competing models

$$I : Y = \alpha + \beta X_1 + \beta_2 X_2 + \varepsilon$$

$$II : Y = \alpha' + \beta' X_3 + \varepsilon$$

[5]

Question 3

The following data were obtained from a calibration check on laboratories A, B and C. Each laboratory was sent samples of standard wire, yielding 13 data points in all. The wire breaking strengths in kilograms are given below:

A	58	63	60	67	
B	70	69	68	70	73
C	61	68	62	53	

Let Y_{ij} denote the breaking strength of the jth wire in laboratory i. Then $\sum y_{ij} = 842$ and $\sum y_{ij}^2 = 54934$.

(a) Construct a one-way analysis of variance table. Test the hyppothesis that the three laboratories each give the same mean breaking strength and report your conclusions. [11]

(b) Prior to running the experiment two contrasts were proposed, one to compare laboratories A and C and the other to compare laboratories A and C together against laboratory B. Write down the two contrasts and show that they are othogonal. [4]

(c) Test the significance of the first contrast in (b) at the 5% level. [4]

(d) Calculate the raw residuals from fitting the one-way analysis of variance model and plot them against laboratory. Comment on your plot. [6]

Question 4

Question 5

An experimenter wanted to compare the effects of two fertilisers on four different types of wheat. She had available 24 plots, each of one third of an acre. She allocated each of the eight combinations of fertiliser and wheat to three plots in a completely randomised design. The yields were as follows, measured in bushels per plot.

SEE NEXT PAGE

	Wheat type				Totals
FertiliserBrand	1	2	3	4	
1	19.4	25.0	24.8	23.1	280.8
	20.6	24.0	26.0	24.3	
	20.0	24.5	25.4	23.7	
2	22.6	25.6	27.6	25.4	302.1
	21.6	26.8	26.4	24.5	
	22.1	26.2	27.0	26.3	
Totals	126.3	152.1	157.2	147.3	582.9

The total sum of squares $CS(y, y) = 117.37$.
Write down the full linear model for such a study, stating clearly the assumptions underlying the model. 5
Give the full analysis of variance table and perform any necessary tests. Interpret your results. Which model is most suitable for these data? 14
Sketch the interaction plot and comment on whether this confirms the results you have found. 6

Obs	Supplier				
	1	2	3	4	5
1	23.46	23.59	23.51	23.28	23.29
2	23.48	23.46	23.64	23.40	23.46
3	23.56	23.42	23.46	23.37	23.37
4	23.39	23.49	23.52	23.46	23.32
5	23.40	23.50	23.49	23.39	23.38
Total	117.29	117.46	117.62	116.9	116.82

[8]

SEE NEXT PAGE

Question 6

In the 1840's and 1850's the Scottish physicist James Forbes was interested in developing a method for estimating altitude on a hillside from measurement of the boiling point of water there. The temperature at which water boils is affected by atmospheric pressure, which in turn is affected by altitude. As part of this study Forbes collected the following data on atmospheric pressure (y measured in degrees Fahrenheit) at 17 locations in Scotland and in the Alps.

y	x	y	x
20.79	194.5	20.79	194.3
22.40	197.9	22.67	198.4
23.15	199.4	23.35	199.4
23.89	200.9	23.99	201.1
24.02	201.4	24.01	201.3
25.14	203.6	26.57	204.6
28.49	209.5	27.76	208.6
29.04	210.7	29.88	211.9
30.66	212.2		

A statistician fitted a series of models to these data. Some extracts from the output are given below.

(a) Model 1 is a simple linear regression model. Write down this model and any necessary assumptions. [2]

(b) On the basis of the output it was decided to try omitting observation 12 to give Model 2. Explain why this was done and how it was achieved in R. [4]

(c) On the basis of the output it was decided to include a quadratic term in x. Explain why this was done. [2]

(d) What hypothesis is the Analysis of Variance Table for Model 3 testing? [2]

(e) Justify Model 3 as a better model than Model 2. [3]

(f) In Model 4 y is transformed to $\log_e y$ (ly). Why is this transformed considered? [2]

(g) Would you use Model 3 or Model 4 for prediction? Justify your answer. [3]

(h) Predict the value of y for another location with $x = 200.2$ using model 4. Explain how you would calculate a 95% prediction interval. (You should give the necessary formula but do not evaluate it.) [4]

(i) Define Cook's distance. How do we use it to determine if any observation has very high influence? [3]

SEE NEXT PAGE

```
> y<-c(20.79,22.40,23.15,23.89,24.02,25.14,28.49,29.04,30.06,20.79,22.67,23.35,23.99,24
> x<-c(194.5,197.9,199.4,200.9,201.4,203.6,209.5,210.7,212.2,194.3,198.4,199.9,201.1,20
> model1.fit<-lm(y~x)
> summary(model1.fit)

Call: lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 *** x
0.52289    0.01011   51.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2328 on 15 degrees of freedom Multiple
R-Squared: 0.9944,     Adjusted R-squared: 0.9941 F-statistic:
2677 on 1 and 15 DF,  p-value: < 2.2e-16

> anova(model1.fit)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 145.125 145.125  2677.1 < 2.2e-16 *** Residuals 15
0.813   0.054
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model1.hat<-lm.influence(model1.fit)$hat
> model1.res<model1.fit$residuals
Error: Object "model1.res" not found
> model1.res<-model1.fit$residuals
> SRES1<-model1.res/sqrt(deviance(model1.fit)*(1-model1.hat)/model1.fit$df)
> FITS1<-model1.fit$fitted
> plot(FITS1,SRES1)
> title("Model 1:Standardised residuls versus fitted values")
> qqnorm(SRES1)
> qqline(SRES1)
>
> x2<-x[c(1:14,16:17)]
> y2<-y[c(1:14,16:17)]
> model2.fit<-lm(y2~x2)
```

SEE NEXT PAGE

```
> x2;y2
 [1] 194.5 197.9 199.4 200.9 201.4 203.6 209.5 210.7 212.2 194.3 198.4 199.9
[13] 201.1 201.3 208.6 211.9
 [1] 20.79 22.40 23.15 23.89 24.02 25.14 28.49 29.04 30.06 20.79 22.67 23.35
[13] 23.99 24.01 27.76 29.88
> summay(model2.fit)
Error: couldn't find function "summay"
> summary(model2.fit)

Call: lm(formula = y2 ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21493 -0.09546 -0.01500  0.09049  0.27793

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.667294   1.419984  -56.81   <2e-16 *** x2
0.520738   0.006997   74.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1608 on 14 degrees of freedom Multiple
R-Squared: 0.9975,     Adjusted R-squared: 0.9973 F-statistic:
5538 on 1 and 14 DF,  p-value: < 2.2e-16

> anova(model2.fit)
Analysis of Variance Table

Response: y2
          Df  Sum Sq Mean Sq F value    Pr(>F)
x2         1 143.150 143.150  5538.2 < 2.2e-16 *** Residuals 14
0.362   0.026
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model2.hat<-lm.influence(model2.fit)$hat
> model2.res<-model2.fit$residuals
> SRES2<-model2.res/sqrt(deviance(model2.fit)*(1-model2.hat)/model2.fit$df)
> FITS2<-model2.fit$fitted
> plot(FITS2,SRES2)
> title("Model 2:Standardised residuls versus fitted values")
> qqnorm(SRES2)
> qqline(SRES2)
>
> xsq=x2*x2
> model3.fit<-lm(y~x2+xsq)
```

SEE NEXT PAGE

```
Error in model.frame(formula, rownames, variables, varnames,
extras, extranames,  :
        variable lengths differ
> model3.fit<-lm(y2~x2+xsq)
> summary(model3.fit
+ )

Call: lm(formula = y2 ~ x2 + xsq)

Residuals:
      Min        1Q    Median        3Q       Max
-0.145936 -0.054387  0.009275  0.054030  0.070662

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.591832  25.079066   4.649 0.000455 *** x2
-1.416466   0.246236  -5.752 6.68e-05 *** xsq           0.004752
0.000604   7.868 2.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06951 on 13 degrees of freedom Multiple
R-Squared: 0.9996,     Adjusted R-squared: 0.9995 F-statistic:
1.485e+04 on 2 and 13 DF,  p-value: < 2.2e-16

> anova(model3.fit)
Analysis of Variance Table

Response: y2
          Df  Sum Sq Mean Sq   F value    Pr(>F)
x2         1 143.150 143.150 29630.638 < 2.2e-16 *** xsq        1
0.299   0.299    61.903 2.682e-06 *** Residuals 13   0.063   0.005
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model3.hat<-lm.influence(model3.fit)$hat
> model3.res<-model3.fit$residuals
> SRES3<-model3.res/sqrt(deviance(model3.fit)*(1-model3.hat)/model3.fit$df)
> FITS3<-model3.fit$fitted
> plot(FITS3,SRES3)
> title("Model 3:Standardised residuls versus fitted values")
> qqnorm(SRES3)
> qqline(SRES3)
> ly<-log(y2)
> model4.fit<-lm(ly~x2)
> summary(model4.fit)
```

SEE NEXT PAGE

```
Call: lm(formula = ly ~ x2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0048082 -0.0014595  0.0004546  0.0020358  0.0031219

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9517662  0.0231021   -41.2 5.16e-16 *** x2
0.0205186  0.0001138   180.2  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002616 on 14 degrees of freedom
Multiple R-Squared: 0.9996,     Adjusted R-squared: 0.9995
F-statistic: 3.249e+04 on 1 and 14 DF,  p-value: < 2.2e-16

> anova(model4.fit)
Analysis of Variance Table

Response: ly
          Df   Sum Sq  Mean Sq F value    Pr(>F)
x2         1 0.222252 0.222252   32485 < 2.2e-16 *** Residuals 14
0.000096 0.000007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model4.hat<-lm.influence(model4.fit)$hat
> model4.res<-model4.fit$residuals
> SRES4<-model4.res/sqrt(deviance(model4.fit)*(1-model4.hat)/model4.fit$df)
> FITS4<-model4.fit$fitted
> plot(FITS4,SRES4)
> title("Model 4:Standardised residuls versus fitted values")
> qqnorm(SRES4)
> qqline(SRES4)
>
```