GX1995

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II (SPRING) EXAMINATIONS, 2003/2004

THIRD ENGINEERING EXAMINATION

Module Code:	**MA338**
Module:	**STATISTICS**
External Examiner	Dr. T. C. Bailey
Internal Examiner	Prof. J. P. Hinde

Instructions:

Duration: **Two** Hours.

Answer any *Three* questions.

All questions, but not necessarily parts therein, carry equal marks.

Relevant tables and formulæ are supplied.

Requirements:

Statistical Tables
Graph Paper

Question 1 is on the next page

1. Consider a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

(a) Which is the response variable and which is the explanatory variable? Are they treated differently? If so, why?

(b) What is the interpretation of the linear model $\beta_0 + \beta_1 x$ and, in particular, the coefficients β_0 and β_1

(c) What assumptions are made about the error terms ϵ_i?

Describe briefly the least-squares principle for estimating β_0 and β_1. What quantity do we minimize?

Writing the fitted values from the model as $\widehat{y}_i$, an important result is that

$$S_{yy} = SS_R + SS_E,$$

where

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SS_R = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 \qquad SS_E = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

Explain the meaning and the importance of these quantities, including how they can be used to form an ANOVA table and how this can be used to test the significance of the regression on x, i.e. to test

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0.$$

Consider the following data:

x	3	4	5	6	7	8	9
y	5	8	8	9	11	10	11

The fitted regression line is

$$3.5 + 0.9\,x$$

Calculate the fitted values ($\widehat{y}_i$) and hence the residuals and SS_E.
Given that $S_{yy} = 26.86$, construct the ANOVA table and test the significance of the regression at the 0.05 level.

Question 2 is on the next page

2. (a) Consider the following MINITAB regression output on vehicle fuel consumption (gallons per 100 miles), selling prices, and vehicle types:

Regression Analysis: gpm100 versus Min.Price, Max.Price, ...

```
The regression equation is
gpm100 = 3.11 + 0.0120 Min.Price + 0.0300 Max.Price + 0.284 Large
         + 0.165 Midsize + 1.37 Van

Predictor        Coef      SE Coef         T        P
Constant       3.1060       0.1358     22.87    0.000
Min.Pric      0.01203      0.01705      0.71    0.484
Max.Pric      0.03002      0.01481         A    0.049
Large          0.2837       0.1551      1.83    0.075
Midsize        0.1651       0.1447      1.14    0.260
Van            1.3735       0.1710      8.03    0.000

S = 0.3393      R-Sq = B%      R-Sq(adj) = 72.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5           C           D     25.87    0.000
Residual Error     E      4.8351      0.1151
Total             47     19.7267

Source       DF      Seq SS
Min.Pric      1      6.2893
Max.Pric      1           F
Large         1      0.0043
Midsize       1      0.1274
Van           1      7.4292
```

i) Find the values of A, B, C, D, E and F.

ii) If the variable `Van` was not included in the regression model, how would this affect the value of SS_E (the residual error sum of squares)?

iii) Use the partial F–test to test the significance of the different vehicle types, i.e.

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

at $\alpha = 0.05$ (where these β's are the coefficients of the variables `Large`, `Midsize`, and `Van`). Interpret your findings.

iv) What can you say about the relationship between the explanatory variable `Min.Price` and the other explanatory variables? Hint: consider the significance of terms in the fitted regression model and the sequential sums of squares.

Question 2 is continued on the next page

(b) Consider the following MINITAB stepwise regression output:

Stepwise Regression:
y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Step	1	2	3	4
Constant	117.57	103.10	71.65	52.58
x4	-0.738	-0.614	-0.237	
T-Value	-4.77	-12.62	-1.37	
P-Value	0.001	0.000	0.205	
x1		1.44	1.45	1.47
T-Value		10.40	12.41	12.10
P-Value		0.000	0.000	0.000
x2			0.416	0.662
T-Value			2.24	14.44
P-Value			0.052	0.000
S	8.96	2.73	2.31	2.41
R-Sq	67.45	97.25	98.23	97.87
R-Sq(adj)	64.50	96.70	97.64	97.44
C-p	138.7	5.5	3.0	2.7

i) Explain the meaning of this output, describe what is happening at each step and give the equation of the final model.

ii) If we had used a forward selection procedure instead of stepwise, what would the final model have been?

iii) Describe, briefly, what we mean by ***best subsets regression.***

Question 3 is on the next page

3. In an investigation on workplace training methods for the time taken to complete a task, 15 workers were divided randomly into three groups. During the training, group A were praised publicly for their work, group B were criticised and group C were ignored. At the end of the training the workers were timed for the completion of the task giving the following data:

Group	Time (mins)				
A	21	14	13	19	15
B	19	28	26	26	19
C	28	30	29	24	27

(a) Perform a one-way analysis of variance for these data; construct the ANOVA table and test, at the $\alpha = 0.05$ level, the null hypothesis of no difference in the population mean times to complete the task.

(b) What proportion of the total variation in completion times is explained by differences between the groups?

(c) Use *Fisher's Least Significant Difference* method to determine which of the three groups are significantly different from one another.

(d) The analysis of variance model is sometimes written as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \text{for } j = 1, \ldots, r,\ i = 1, \ldots, k.$$

where μ is the overall population mean. Interpret each of the other terms in this model and explain how your above analyses relate to this model.

Question 4 is on the next page

4. (a) An important aspect of any regression analysis is the checking of model assumptions. What are the assumptions that we need to check, both about the systematic part of the model and the error terms? Explain how the residuals can be used to check each of these specific aspects. [You may find the plots in Figure 1 a useful reminder.]

Now consider the output in Figure 1 in more detail. What do each of these various plots tell you about the adequacy of the regression model in this situation? How might the model be changed to try and improve matters?

(b) For each regression model fit Minitab also prints out a list such as follows:

Unusual Observations

Obs	Body Wt	Brain Wt	Fit	SE Fit	Residual	St Resid
19	2547	4603.2	2552.6	119.7	2050.5	6.56RX
32	62	1320.0	150.9	43.0	1169.1	3.52R
33	6654	5711.9	6522.0	310.6	-810.1	-6.49RX

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Explain what it means for a point to have ***large influence***, using a diagram for illustration. Why might we be particularly worried about points which have both a large standardized residual and a large influence?

(c) In a simple linear regression model (with a single explanatory variable) the fitted value at an arbitrary point x^* is $\hat{\beta}_0 + \hat{\beta}_1 x^*$ with variance

$$\sigma^2 \left\{ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right\}$$

where $\bar{x}$ and S_{xx} are the mean and sum-of-squares of the x-values used in fitting the regression model. Explain how we can use these to obtain a $100(1-\alpha)\%$ confidence interval for the population mean response at x^*. What does the width of this confidence interval depend on? Where is it narrowest in terms of the value of x^*?

When we are interested in making predictions for a future y-value at x^*, why is the prediction interval wider? That is, how does the prediction variance differ from that for the population mean response?

Use the following results to check the calculation for the 95% confidence interval and to obtain the 95% prediction interval.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	3.864	1.024	(1.826, 5.902)	(??? , ???)

S = 4.904 on 80 degrees of freedom

Figure 1 is on the next page

Figure 1: Residual Diagnostic Plots