Gx 2091

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II (SPRING) EXAMINATIONS, 2002/2003

THIRD ENGINEERING EXAMINATION

Module Code:	**MA338**
Module:	**STATISTICS**
External Examiner	Dr. D. Harrington
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

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

Interpret each of the elements of this model, including reference to:

- the role of the variables y and x
- the linear model $\beta_0 + \beta_1 x$ and the coefficients β_0 and β_1
- the error terms ϵ_i; what assumptions are made about these?

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	13	12	12	19	17	20	21

The fitted regression line is

$$6.64 + 1.61\,x$$

Calculate the fitted values ($\widehat{y}_i$) and hence the residuals and SS_E.
Given that $S_{yy} = 91.43$, 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 the quarter-mile time for a sample of performance cars:

Regression Analysis: QMT versus DISP, HP, CB, WT, CYL

The regression equation is
QMT = 18.1 - 0.0125 DISP - 0.00118 HP - 0.613 CB + 0.00213 WT - 0.381 CYL

Predictor	Coef	SE Coef	T	P
Constant	18.1024	0.9323	19.42	0.000
DISP	-0.012498	0.004952	-2.52	0.018
HP	-0.001181	0.006948	-0.17	0.866
CB	-0.6126	0.1995	A	0.005
WT	0.0021283	0.0004018	5.30	0.000
CYL	-0.3811	0.2022	-1.88	0.071

S = 0.8576 R-Sq = B % R-Sq(adj) = 76.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	75.781	C	D	0.000
Residual Error	E	18.386	0.735		
Total	30	94.166			

Source	DF	Seq SS
DISP	1	16.087
HP	1	35.160
CB	1	0.449
WT	1	F
CYL	1	2.612

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

ii) If the variable CYL 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

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

at $\alpha = 0.05$ (where these β's are the coefficients of the variables CB, WT and CYL). Interpret your findings.

iv) What can you say about the relationship between the explanatory variable HP 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:
QMT versus S, CYL, T, G, DISP, HP, CB, DRAT, WT

Step	1	2	3	4	5
Constant	16.69	20.01	21.60	17.17	16.02
S	2.59	2.93	2.05	2.05	2.00
T-Value	5.80	7.50	4.01	4.72	4.68
P-Value	0.000	0.000	0.000	0.000	0.000
G		-0.93	-0.91	-0.23	
T-Value		-3.52	-3.71	-0.82	
P-Value		0.002	0.001	0.421	
HP			-0.0088	-0.0165	-0.0178
T-Value			-2.41	-4.30	-5.19
P-Value			0.023	0.000	0.000
WT				0.00094	0.00110
T-Value				3.39	5.51
P-Value				0.002	0.000
S	1.23	1.04	0.960	0.814	0.809
R-Sq	53.67	67.88	73.58	81.69	81.22
R-Sq(adj)	52.08	65.59	70.65	78.87	79.13
C-p	40.8	22.0	15.7	5.8	4.5

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 the tensile strength of three different rubber compounds, four samples of each compound were subjected to a strength test giving the following data:

Compound	Strength			
A	32	30	33	32
B	32	34	33	34
C	35	36	36	35

(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 strengths of the three compounds.

(b) What proportion of the total variation in recorded strengths is explained by differences between the compounds?

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

(d) In fact, the three compounds were made by using different amounts of a special additive at rates of 10, 20, and 30 g/Kg for the compounds A, B, and C, respectively. How would you explore the relationship between the additive and the resulting strength of the rubber using a regression model? *You do not need to do the analysis.* Sketch a graph indicating the relationship between the fitted means from the regression and the one-way analysis of variance.
Fitting a regression model results in a regression sum of squares of 28.1; what do you conclude?

Question 4 is on the next page

4. This question concerns the Minitab output on the following two pages.

Horse mussels (*Atrinia*) were sampled from the Marlborough Sounds, New Zealand. The response variable is the mussels' Muscle Mass (M) in grams, the edible portion of the mussel. Measurements were also made on the dimensions of the shell; the shell height (H) and the shell width (W), both measured in mm.

(a) Consider the plots of mass against each of the two recorded dimensions. Discuss the appearance of these plots. Do they suggest that there are likely to be any problems when conducting ordinary multiple regression on these data?

(b) Discuss the output of the first regression fully. Ensure that you make reference to appropriate hypothesis tests. What can you conclude from the diagnostic plots of the residuals? What features of the model are we assessing in these?

(c) Finally, discuss the remaining output. Why do you think that a cube-root transformation has been used? Compare the model with the transformed response to the original regression fit. Again, be sure to make careful reference to the residual plots. Draw conclusions as to your preferred model.

Question 4 is continued on the next page