Gr 2038

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II (SUMMER) EXAMINATIONS, 2003/2004

THIRD ARTS AND SCIENCE 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) (i) Which is the response variable and which is the explanatory variable? Are they treated differently? If so, why?

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

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

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

(c) 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

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

(d) Consider the following data:

x	1	2	3	4	5	6	7
y	5	9	8	9	11	10	13

The fitted regression line is

$$\hat{y} = 5.1 + 1.0\,x$$

Calculate the fitted values ($\widehat{y}_i$) and hence the residuals and SS_E.
Given that $S_{yy} = 37.43$, 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 edited MINITAB output relating to a multiple regression analysis of `MPG.highway` (measured fuel consumption for highway driving) versus five possible explanatory variables (`Range.Price`, `Price`, `Small`, `Large` and `Van`).

```
The regression equation is
MPG.highway = 32.0 - 0.005 Range.Price - 0.185 Price + 3.77 Small
            - 0.730 Large - 7.16 Van
```

Predictor	Coef	SE Coef	T	P
Constant	31.957	1.172	27.27	0.000
Range.Pr	-0.0047	0.1107	-0.04	0.967
Price	-0.18463	0.05699	-3.24	0.002
Small	3.772	1.164	A	0.002
Large	-0.7305	0.9903	-0.74	0.465
Van	-7.159	1.228	-5.83	0.000

S = 2.487 R-Sq = B % R-Sq(adj) = 64.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	550.12	110.02	C	0.000
Residual Error	42	259.86	D		
Total	E	809.98			

Source	DF	Seq SS
Range.Pr	1	12.90
Price	1	208.42
Small	1	F
Large	1	0.37
Van	1	210.26

(i) Find the values of A, B, C, D, E and F. (Note that it may not be possible to find the values in that order.)

(ii) If the variable `VAN` was excluded from the above regression how would this affect the value of SS_R, the regression 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 `Small`, `Large`, and `Van`). Interpret your findings.

(iv) Interpret the meaning of the significant coefficients in the fitted regression model. Note that `Small`, `Large` and `Van` are indicator variables for vehicles of these types.

Question 2 is continued on the next page

(b) Consider the following MINITAB ***best subsets regression*** output for the response variable `MPG.highway` (measured fuel consumption for highway driving) on seven potential explanatory variables (`Min.Price`. `Max.Price`. `Large`, `Midsize`, `Small`, `Sporty` and `Van`).

Vars	R-Sq	R-Sq(adj)	C-p	S	Min. Price	Max. Price	Large	Midsize	Small	Sporty	Van
1	33.0	31.5	44.6	3.4347				X			
2	58.4	56.6	12.9	2.7348	X						X
3	67.0	64.8	3.6	2.4631	X			X			X
4	68.1	65.2	4.2	2.4507	X			X	X		X
5	68.9	65.2	5.1	2.4476	X		X	X	X		X
6	69.6	65.2	6.1	2.4491	X		X	X	X		X
7	69.8	64.5	8.0	2.4749	X	X	X	X	X		X

(i) Explain the meaning of this output.

(ii) Plot R^2 against the number of variables in the model and use this to select a possible model for these data.

(iii) Using C_p which model would you choose? Give reasons for your choice.

Question 3 is on the next page

3. The times require by three workers to perform an assembly-line task were recorded on five randomly selected occasions. The times, to the nearest minute, are given in the following table:

Worker	Time (mins)				
Bertie	9	11	10	12	11
Mary	8	9	9	8	8
Liam	10	9	10	11	9

(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 between the workers times to complete this task.

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

(c) Use *Fisher's Least Significant Difference* method to determine which of the three workers 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 predicted value for a future y-value at a new point x^* is $\hat{\beta}_0 + \hat{\beta}_1 x^*$ with variance

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

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 to obtain a $100(1-\alpha)\%$ prediction interval for the response at x^*. What does the width of this prediction interval depend on? Where is it narrowest in terms of the value of x^*?

How does this differ from deriving a confidence interval for the population mean response at x^*? That is, what is the estimated mean response and how does the variance for this differ from that for a predicted new value?

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

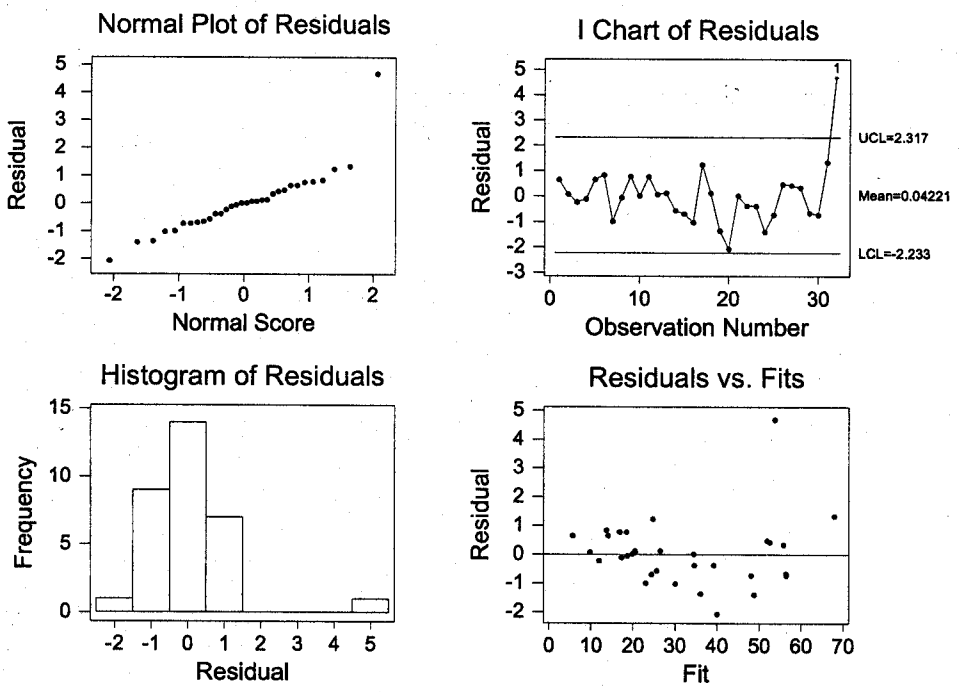
Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
%1	3.864	1.024	(??? , ???)	(-6.105, 13.833)

S = 4.904 on 80 degrees of freedom

Figure 1 is on the next page

Figure 1 : Residual Diagnostic Plots



Question 5 is on the next page

5. Explain how indicator variables can be used to incorporate categorical variables (factors) into regression models for a response variable Y. For a 4-level factor, A, how many indicator variables would there be in the regression model?
What is the relationship between the indicator variable regression model and a one-way analysis of variance of Y over the factor A?

The following MINITAB analyses are on data from a study of birthweight. Birthweights (`wt`) were recorded for 12 male and 12 female babies along with their estimated gestational ages (`age`), from 35 to 42 weeks. The other variable in the analyses is an indicator variable for the male babies (i.e. takes the value 1 if the baby is male and 0 if female).

(a)
```
The regression equation is
wt = 2911 + 113 male

Predictor       Coef     SE Coef        T        P
Constant     2911.33       81.50    35.72    0.000
male           112.7       115.3     0.98    0.339

S = 282.3       R-Sq = 4.2%

Analysis of Variance

Source           DF          SS          MS        F        P
Regression        1       76163       76163     0.96    0.339
Residual Error   22     1753711       79714
Total            23     1829873
```

(i) What is the fitted value predicted from the model for a male baby and what is that predicted for a female baby? Interpret the difference in these fitted values.

(ii) Test (at $\alpha = 0.05$) the difference in population mean birthweight between male and female babies. How does your test relate to the usual two-sample t-test that could be used here?

Question 5 is continued on the next page

(b) The following output, with some missing entries, is from a simple analysis of covariance.

```
The regression equation is
wt = - 1773 + 163 male + 121 age

Predictor        Coef      SE Coef          T        P
Constant      -1773.3        794.6      -2.23    0.037
male           163.04        72.81       2.24    ?????
age            120.89        20.46       5.91    0.000

S = 177.1       R-Sq = 64.0%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         2     1171103      585551        ???     ????
Residual Error    21      658771       31370
Total             23     1829873
```

(i) Write down the separate fitted models for male and female babies.
Sketch these models and interpret each of the parameter estimates in the fitted regression equation. Do you have any reservations about the model?

(ii) Perform an F-test at $\alpha = 0.05$ to test the overall significance of the regression model. What is the interpretation of the null hypothesis for this test?

(iii) Test (at $\alpha = 0.05$) the effect of gender in this model. Explain why your conclusions here are different from those in Section (a).

Formulæ

Partial F-test

$$F = \frac{SS_R\left(\beta_p, \ldots \beta_k \mid \beta_{k-1}, \ldots \beta_2, \beta_1, \beta_0\right) / (p-k+1)}{MS_E}$$

Analysis of Variance

$$\text{Sum of Squares of Total} = SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{\left[\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right]^2}{N}$$

$$\text{Sum of Squares of Treatments} = SS_{\text{treat}} = n\sum_{i=1}^{a}(\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{a}\frac{y_{i\cdot}^2}{n} - \frac{\left[\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right]^2}{N}$$

$$\text{where } y_{i\cdot} = \sum_{j=1}^{n} y_{ij}$$

$$\text{Sum of Squares of Error} = SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

$$SS_E = SS_T - SS_{\text{treat}}$$

Fisher's Least Significant Difference

$$LSD = t_{a(n-1),\frac{\alpha}{2}}\sqrt{\frac{2MS_E}{n}}$$