970

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II (SUMMER) EXAMINATIONS, 2004/2005

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:

Graph Paper

Question 1 is on the next page

1. (a) The relationship between two variables x and y is often summarised by a **correlation coefficient** ρ. Write brief responses to the following questions, using pictures to illustrate your answers where appropriate.

   (i) What range of values can ρ take?

   (ii) If there is a strong positive correlation between x and y, what does this mean? Are there situations where a large value of the sample correlation coefficient can be misleading?

   (iii) If there is a zero correlation between x and y does this mean that the variables are unrelated?

   (iv) What should we do to help us avoid making incorrect inferences from a sample correlation coefficient?

   (b) Consider a simple linear regression model

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

   (i) Which is the response variable and which is the explanatory variable? How does the situation here differ from using correlation?

   (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?

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

   (v) 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 of these quantities.
   For a particular regression analysis on 32 observations, we have $S_{yy} = 140$ and $SS_E = 120$. Use these quantities to construct the ANOVA table and test (at the 0.05 level) the significance of the regression on x, i.e. test

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

Question 2 is on the next page

2. (a) Consider the following edited MINITAB regression output on the forced expiratory volume (FEV) of a sample of men and its possible relationship to cigarette consumption, age, weight, and height.

Regression Analysis: FEV versus CIGS, AGE, HEIGHT, WEIGHT

```
The regression equation is
FEV = 0.16 + 0.00765 CIGS - 0.0319 AGE + 0.0164 HEIGHT + 0.0149 WEIGHT

Predictor        Coef   SE Coef       T      P
Constant        0.156     1.647    0.09  0.925
CIGS         0.007647  0.007105    1.08  0.284
AGE         -0.031910  0.008748   -3.65  0.000
HEIGHT       0.016407  0.009462       A  0.086
WEIGHT       0.014945  0.005447    2.74  0.007

S = 0.495613   R-Sq = B%   R-Sq(adj) = 25.8%

Analysis of Variance

Source           DF       SS       MS      F      P
Regression        4        C   2.3804      D  0.000
Residual Error    E  23.5807   0.2456
Total           100  33.1023

Source  DF  Seq SS
CIGS     1  1.1262
AGE      1  4.3999
HEIGHT   1  2.1461
WEIGHT   1       F
```

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

(ii) What can you say about the effect of cigarette consumption on FEV? Is this a significant effect?

(iii) Use the partial F–test to test the significance (at $\alpha = 0.05$) of weight and height taken together, i.e.

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

where these β's are the coefficients of the variables HEIGHT and WEIGHT. Interpret your findings.

(iv) What can you say about the relationship between the explanatory variables WEIGHT and HEIGHT? Hint: consider the significance of the HEIGHT term in the fitted regression model and the sequential sums of squares.

Question 2 is continued on the next page

(b) Consider now the following MINITAB best subsets regression output for the male FEV data:

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	AGE	WEIGHT	CIGS	HEIGHT
1	15.0	14.1	17.6	0.53324	X			
1	12.2	11.3	21.3	0.54179		X		
2	25.2	23.7	5.8	0.50259	X	X		
2	22.3	20.7	9.7	0.51221	X			X
3	27.9	25.7	4.2	0.49602	X	X		X
3	26.5	24.3	6.0	0.50071	X	X	X	
4	28.8	25.8	5.0	0.49561	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 the other criteria given here (R^2-adjusted and C_p) which models would you select. Why?

(iv) Having used this output to select one, or more, models, what would you do next?

Question 3 is on the next page

3. In a study of the wood reserves in a forest the aim was to find a predictive relationship for the volume of wood (only measurable by destructive methods) in terms of the diameter at a specified height on the trunk . Data was obtained for a sample of 31 trees and a simple linear regression model was fitted giving the following output and fitted line plot:

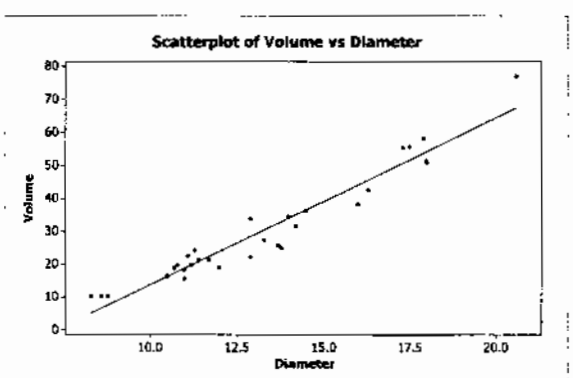
```
The regression equation is
Volume = - 36.9 + 5.07 Diameter

Predictor      Coef  SE Coef       T      P
Constant   -36.943    3.365  -10.98  0.000
Diameter    5.0659   0.2474   20.48  0.000

S = 4.25199   R-Sq = 93.5%   R-Sq(adj) = 93.3%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  7581.8  7581.8  419.36  0.000
Residual Error   29   524.3    18.1
Total            30  8106.1
```



(a) Write down the fitted regression equation and interpret the coefficients. Do these make sense?

(b) Test the significance of the regression and give an approximate 95% confidence interval for the estimated coefficient for diameter. Comment on how well the fitted model explains the data and whether you think that it might be useful for prediction.

(c) Using this model to predict volume for a tree of diameter 20 units gave the following output

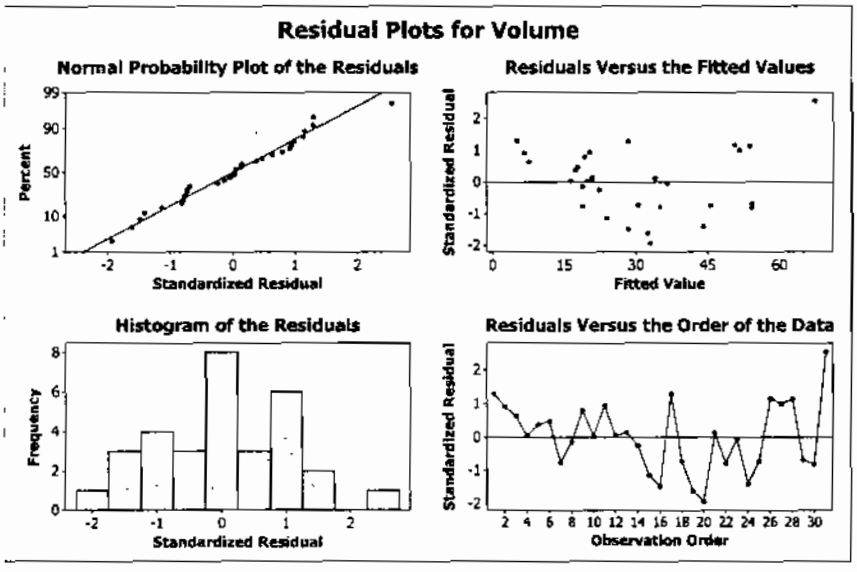
```
New Obs      Fit  SE Fit        95% CI             95% PI
  20      64.374   1.837  (60.618, 68.130)  (54.901, 73.846)
```

Explain the meaning of this output and, in particular, the difference between the 95% CI and 95% PI.

Question 3 is continued on the next page

(d) 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 the figure below a useful reminder.]

(e) Now consider the output below 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?



Question 4 is on the next page

4. (a) In studying the effects of a number of different treatments, what do we mean by a

   - completely randomised design?
   - randomised block design?

   What is the aim of ***blocking***?

   The analysis of variance model for a randomised block design with t treatments and b blocks is sometimes written as

$$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}, \qquad \text{for } i = 1, \ldots, t \; ; \; j = 1, \ldots, b \; ; \; k = 1, \ldots, r$$

   where μ is the overall mean. Interpret each of the other terms in this model and state the usual assumptions that are made.

   (b) To compare the compressive strengths of concrete using three different drying methods (treatments), concrete was mixed in batches that were just large enough to produce three cylinders, one for each drying method. These were then subjected to strength tests. This was repeated 5 times giving the data below.

	Batch				
Treatment	B_1	B_2	B_3	B_4	B_5
T_1	52	47	44	51	42
T_2	60	55	49	52	43
T_3	56	48	45	44	38

   (i) A simple one-way analysis of variance comparing the three treatments gives the following output:

```
Source      DF     SS     MS     F      P
treatment    2   89.2   44.6  1.30  0.307
Error       12  410.4   34.2
Total       14  499.6

S = 5.848   R-Sq = 17.85%
```

   Explain the meaning of this output. Write down the null and alternative hypotheses that are being tested by this analysis. What do you conclude here?

   Why may this not be an appropriate analysis?

Question 4 is continued on the next page

(ii) Now consider the following output from a two-way analysis of variance:

Source	DF	SS	MS	F	P
batch	4	363.600	90.900	15.54	0.001
treatment	2	89.200	44.600	7.62	0.014
Error	8	46.800	5.850		
Total	14	499.600			

S = 2.41868 R-Sq = 90.63%

Once again explain the meaning of the output, write down the hypotheses being tested and give your conclusions. What can you say about the effectiveness of the experimental design? Are there any shortcomings in the experiment as performed?

(iii) To compare the three different drying methods we can look at comparisons between the treatment means allowing for batch variation as below.

```
Bonferroni 95.0% Simultaneous Confidence Intervals

treatment = 1  subtracted from:

treatment   Lower  Center  Upper  ------+---------+---------+---------
2          -0.013   4.600  9.213                  (-------*-------)
3          -5.613  -1.000  3.613          (-------*-------)
                                  ------+---------+---------+---------
                                      -6.0       0.0       6.0
treatment = 2  subtracted from:

treatment   Lower  Center   Upper  ------+---------+---------+---------
3          -10.21  -5.600  -0.9868 (-------*-------)
                                  ------+---------+---------+---------
                                      -6.0       0.0       6.0
```

- What do you conclude about the three treatments?
- Why have we used the Bonferroni method to make these comparisons rather than simple 95% confidence intervals?
- What has been the effect on these comparisons of allowing for batch variation?

Question 5 is on the next page

5. (a) 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?

(b) The following MINITAB analyses are on data from a study of womens cholesterol levels based on samples from two US states, Iowa and Nebraska. The age of the women was also recorded. The `Iowa` and `Nebraska` variables are indicator variables for the two states.

Regression Analysis: chol versus Iowa, Nebraska

```
* Nebraska is highly correlated with other X variables
* Nebraska has been removed from the equation


The regression equation is
chol = 217 - 9.4 Iowa
```

Predictor	Coef	SE Coef	T	P
Constant	217.11	13.91	15.61	0.000
Iowa	-9.38	22.97	?????	0.686

S = 60.63 R-Sq = 0.6% R-Sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	613	613	????	0.686
Residual Error	28	102924	3676		
Total	29	103537			

(i) Obtain the missing F-value and comment on the size of the p-value.

(ii) Calculate the fitted values for the two groups. What do these fitted values correspond to?

(iii) Use the output to perform a t-test (at $\alpha = 0.05$) of the difference between the population mean cholesterol levels of women in the two states. 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

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

Regression Analysis: chol versus age, Iowa, Nebraska

```
* Nebraska is highly correlated with other X variables
* Nebraska has been removed from the equation


The regression equation is
chol = 82.2 + 3.04 age - 35.9 Iowa
```

Predictor	Coef	SE Coef	T	P
Constant	82.16	24.47	3.36	0.002
age	3.0414	0.5102	5.96	0.000
Iowa	-35.91	16.00	?????	?????

S = 40.57 R-Sq = 57.1% R-Sq(adj) = 53.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	59103	29551	?????	?????
Residual Error	27	44434	1646		
Total	29	103537			

(i) Write down the separate fitted models for the two groups.
Sketch these fitted models and interpret each of the parameter estimates.

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

(iii) Use a t-test (at $\alpha = 0.05$) to test the difference between states in this model. Comment on your result and compare it and the estimates for Iowa with those from the first analysis. Explain any similarities and differences. Can you think of any reason for what you have observed?