GX1984

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER I EXAMINATIONS 2002-2003

MA 322 – APPLIED STATISTICS

Dr. T.C. Bailey
Dr. J. Newell

Time allowed: *Two* hours

Answer question 1[30 marks] and any 2 of questions 2-4 [20 marks each].

Question 1.

[3 marks for each question]

(a) When asked to state the simple linear regression model, a student wrote it is as follows: $E\{Y_i\} = \beta_0 + \beta_1 X_i + \varepsilon_i$. Do you agree?

(b) For a simple linear regression the estimate of the slope is 5 with an estimated standard error of 2.5. Assume that the sample size is 24. Give a 95% confidence interval for your estimate of the true slope.

(c) The correlation between two variables X and Y is 0.80. What percent of the variation in Y can be explained by X using a simple linear regression?

(d) Write down the line of best fit between X and Y given that the estimate of the slope is 1.5, $\bar{Y} = 20$ and $\bar{X} = 10$. Use this line to provide an estimate of Y when X is equal to 30.

(e) What can you conclude regarding the need for an intercept in a regression model if you compute a 95% confidence interval for β_0 as [-1.0, 1.0]?

(f) A client wants to use a linear regression to predict a variable Y using a single predictor variable X. The residual plots indicate that the linearity assumption is not valid. What advice would you give to this client?

(g) In a multiple regression the standardised residual for a particular observation is 0.05. Would you consider this observation to be an outlier? Explain why or why not.

(h) A multiple regression is fitted with response variable Y and explanatory variables X_1, X_2 and X_3. There are 244 cases. What are the degrees of freedom for (a) SSR (b) SSE and (c) SSTO?

(i) In a simple linear regression two intervals are reported for the value of the response variable when the explanatory variable has the value X = 6: [83, 89] and [76, 96]. Which is the prediction interval and which is the confidence interval for the mean response and what is the estimate of E{Y} given X is equal to 6?

(j) In an analysis of the model $Y = \beta_0 + \beta_1 X + \varepsilon$, the estimate of β_1 was 5.2 with an estimated standard error of 1.4. What is the value of the t statistic for testing the null hypothesis that $\beta_1 = 0$?

Question 2.

(a)

(i) What are the underlying assumptions relating to the Simple Linear Regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$? *[3 Marks]*

(ii) What is the meaning of the error term in the Simple Linear Regression model? *[2 Marks]*

(iii) What is the difference between the error term and the residuals? *[1 Mark]*

(b)

Data were collected from several publishing companies in order to investigate the relationship between the percentage of expenditure, E, a publishing house spends on advertising and the change in revenue, R (expressed as a percentage) at the end of the following year. A statistician fitted a simple linear regression model and decided a transformation of the response variable was needed on the basis of suitable residual plots. She decided that the best transformation (using the Box-Cox procedure) was to raise the response variable Y to the power 0.25 (i.e. $Y^{0.25}$). A scatterplot and some relevant Minitab output are given below.

(i) What is the sample size? *[1 Mark]*

(ii) On the basis of Figure 1 explain briefly why you think a transformation was necessary. *[2 Marks]*

Question 2 continued.

(iii) Does it look like the transformation was worthwhile? Make specific reference to the residual plots included. *[2 Marks]*

(iv) Give the fitted regression equation for the model where Revenue is predicted by Expenditure. *[1 Mark]*

(v) What percent of the variation in the response is explained by the model where Revenue is predicted by Expenditure? *[1 Mark]*

(vi) Carry out a suitable significance test (null and alternative hypotheses, test statistic, degrees of freedom, P-value, and conclusion) for the regression coefficient (i.e. the slope) in the model where Revenue (transformed) is predicted by Expenditure. *[4 Marks]*

(vii) What Revenue would you predict for a particular firm that plans to spend 10% on advertising in the future? Your answer should include a suitable interval estimate. *[3 Marks]*

Figure 1.
Scatterplot of Revenue against Expenditure.

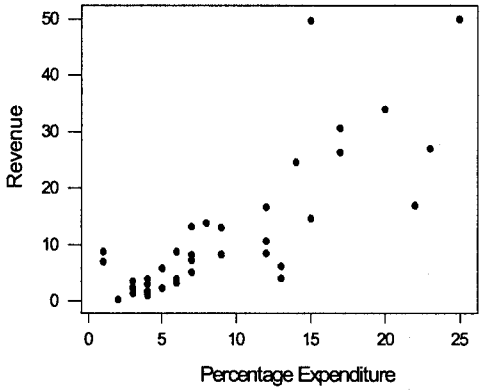
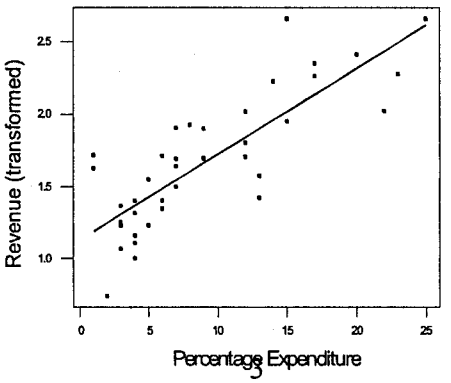
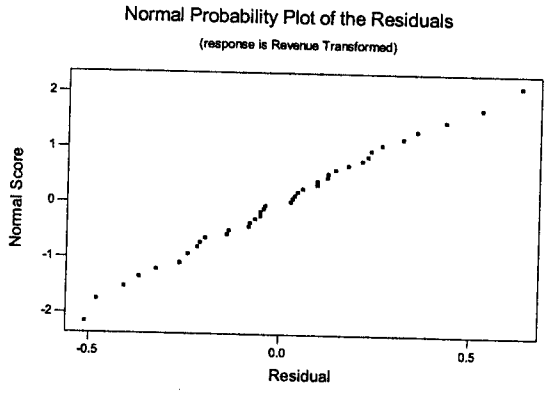
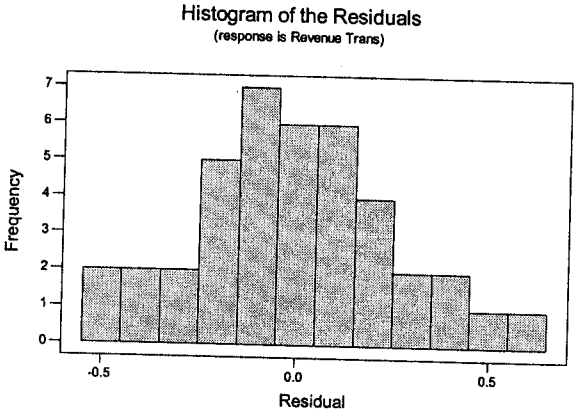


Figure2.
Scatterplot of Revenue (transformed) against Expenditure.



Minitab Output for Question 2.





The regression equation is
Revenue Trans = 1.13 + 0.0593 Expenditure

Predictor	Coef	SE Coef	T	P
Constant	1.13285	0.07384	15.34	0.000
Expenditure	0.059284	0.006724	8.82	0.000

S = 0.2676 R-Sq = 67.2% R-Sq(adj) = 66.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5.5672	5.5672	77.75	0.000
Residual Error	38	2.7211	0.0716		
Total	39	8.2883			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	1.7257	0.0428	(1.6390, 1.8124)	(1.1771, 2.2743)

Values of Predictors for New Observations

New Obs	P
1	10.0

Question 3.

a) The general linear model can be formulated in matrix terms as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a vector of responses, $\boldsymbol{\beta}$ is a vector of parameters, $\mathbf{X}$ is a matrix of coefficients and $\boldsymbol{\varepsilon}$ is a vector of independent normal variables with mean 0 and variance-covariance matrix $\sigma^2\mathbf{I}$.

What is the least-square estimator $\mathbf{b}$ of $\boldsymbol{\beta}$ in matrix notation and what information is provided by the formula $\hat{V}(\mathbf{b}) = (X'X)^{-1}s^2$? [2 Marks]

b) Data relating to salaries in a small American university were collected for 52 lecturers. The data were collected as evidence in a class-action case to judge a claim of salary discrimination against women.

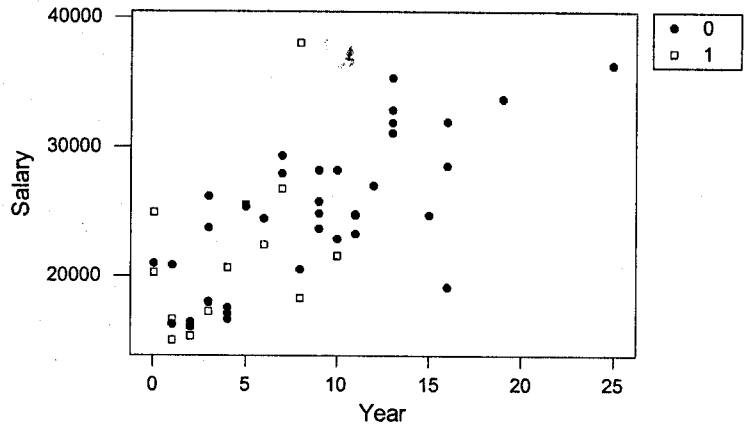
Variables on which data were collected were as follows:

```
C1:   Salary      (in dollars)
C2:   Year        Years in current rank
C3:   Sex         (coded as 1 if female, 0 if male)
```

(i) Based on the graph below and the output overleaf what can you say regarding evidence of salary discrimination against women? Make specific reference to the evidence for and against the claim contained in both models. [14 Marks]

(ii) What relevant information is provided for you in the residuals in this analysis with respect to possible discrimination? [4 Marks]

Minitab Output for Question 3.



Model 1

The regression equation is
Salary = 18065 + 759 Year + 201 Sex

Predictor	Coef	SE Coef	T	P
Constant	18065	1248	14.48	0.000
Year	759.0	118.3	6.41	0.000
Sex	201	1455	0.14	0.890

S = 4306 R-Sq = 49.1% R-Sq(adj) = 47.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	877036388	438518194	23.65	0.000
Residual Error	49	908693470	18544765		
Total	51	1785729858			

Source	DF	Seq SS
Year	1	876680907
Sex	1	355481

Unusual Observations

Obs	Year	Salary	Fit	SE Fit	Residual	St Resid
1	25.0	36350	37041	2047	-691	-0.18 X
21	16.0	19175	30210	1108	-11035	-2.65R
24	8.0	38045	24339	1241	13706	3.32R

Model 2

The regression equation is
Salary = 18223 + 741 Year - 571 Sex + 169 Year*Sex

Predictor	Coef	SE Coef	T	P
Constant	18223	1309	13.92	0.000
Year	741.0	126.2	5.87	0.000
Sex	-571	2297	-0.25	0.805
Year*Sex	169.1	387.0	0.44	0.664

S = 4342 R-Sq = 49.3% R-Sq(adj) = 46.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	880635387	293545129	15.57	0.000
Residual Error	48	905094471	18856135		
Total	51	1785729858			

Source	DF	Seq SS
Year	1	876680907
Sex	1	355481
Year*Sex	1	3599000

Unusual Observations

Obs	Year	Salary	Fit	SE Fit	Residual	St Resid
1	25.0	36350	36748	2170	-398	-0.11 X
21	16.0	19175	30079	1156	-10904	-2.61R
24	8.0	38045	24932	1847	13113	3.34R
31	10.0	21600	26753	2460	-5153	-1.44 X

Question 4.

A financial analyst responsible for cost reduction at a production plant is interested in determining what factors affect the amount of water used at the plant. She decided to investigate water usage by collecting 17 observations on the plant's water usage (Y) and four other variables, namely:

X_1 = average monthly temperature (^oF)

X_2 = average production (€)

X_3 = number of plant operating days in the month

and

X_4 = numbers of persons on the monthly plant payroll

A multiple linear regression model was fitted using all four predictors.

(i) Use the matrix scatterplot and correlation matrix to carry out a preliminary assessment of the relationship between water usage and each of the 4 predictors.

[4 Marks]

(ii) Does a multiple linear regression model appear appropriate based on the matrix scatterplot?

[2 Marks]

(iii) Is there evidence of multicollinearity present amongst some of the variables?

[2 Marks]

(iv) Decide, based on the output below, which model you think is the most appropriate and give clear reasons for your decision.

[8 Marks]

(v) What additional information would you need to complete this multiple regression analysis for the financial analyst in question?

[2 Marks]

(vi) Summarise your findings for the financial analyst in question.

[2 Marks]