GX2086

OLLSCOIL NA hÉIREANN, GAILLIMH
NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER I EXAMINATIONS 2002-2003

MA 322 – APPLIED STATISTICS

Dr. D. Harrington
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]

(i) If you compute a 99% confidence interval for a population mean as 112.5 to 118.4, what can you conclude?

(ii) For a simple linear regression the estimate of the slope is -2 with an estimated standard error of 2. Assume that the sample size is 36. Give a 95% confidence interval for your estimate of the true slope.

(iii) A simple linear regression model with a transformed response variable $\sqrt{Y} = \beta_0 + \beta_1 X + \varepsilon$ fitted to a set of data gave b_0=6 and b_1=3. Find the predicted value of the response variable and the residual when the explanatory variable is equal to 10 and the response equal to 5.5.

(iv) A simple linear regression fitted to a set of data resulted in an R^2 value of 0.64. What is the correlation between Y and X?

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

(vi) A study found a correlation of r = -0.61 between the sex of a worker and his or her income. What can you conclude?

(vii) What patterns in a residual plot suggest a) a non-linear relationship and b) homogeneity of variance?

(viii) A multiple regression is run with response variable Y and explanatory variables X_1 and X_2. There are 100 cases. What are the degrees of freedom for (a) SSR, (b) SSE and (c) SSTO?

(ix) In a regression model how many indicators are needed to represent a nominal (categorical) variable with k categories?

(x) Which of the following models are acceptable according to the C_P criterion: (a) five explanatory variables with C_P =1.8, (b) two explanatory variables with C_P =4.6, (c) eight explanatory variables with C_P =8.9?

Question 2.

(a)

(i) Write down the normal errors Simple Linear Regression model and clearly define each term.

[3 Marks]

(ii) What are the underlying assumptions relating to this model?

[2 Marks]

(b)

Data were collected on the fuel efficiency of a particular new people carrier. The response variable Y is miles per gallon (mpg), a measure of fuel efficiency, and the explanatory variable X is the average speed in miles per hour (mph). The appropriate Minitab output for the analysis of these data in addition to a scatterplot of the data (with the line of best fit superimposed) and some residual plots are given below.

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

(ii) Give the fitted regression equation for the model where mpg is predicted by mph and describe the relationship in words. *[1 Mark]*

(iii) What percent of the variation in Y is explained by its relationship with X? *[1 Mark]*

(iv) Give the results of the significance test (null and alternative hypotheses, test statistic, degrees of freedom, P-value, and conclusion) for the regression coefficient of mph in the model where mpg is predicted by mph. *[6 Marks]*

(v) The specification of the model for the analysis (where mpg is predicted by mph) includes a term for the variance σ^2. What is the value of the estimate of this parameter? *[2 Marks]*

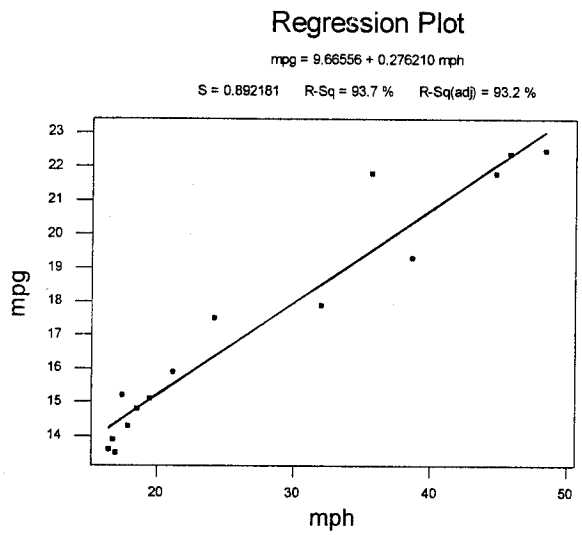
(vi) What is the predicted miles per gallon at 40 miles per hour? *[1 Marks]*

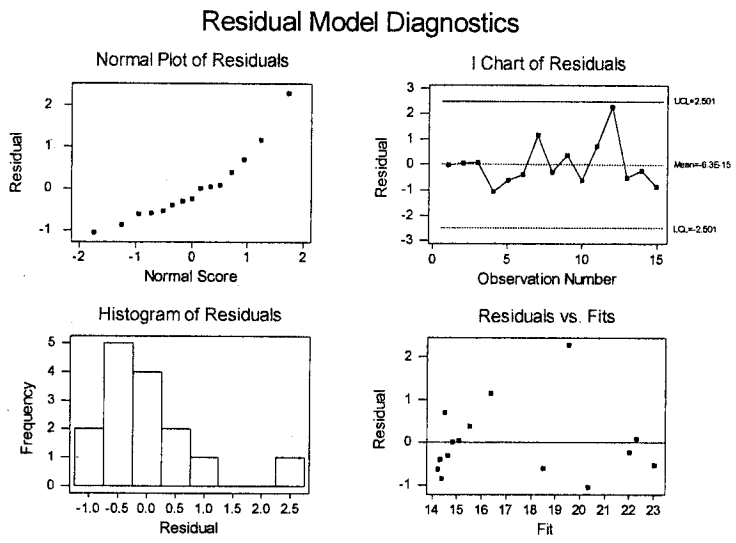
(vii) Summarize information from the residuals plots that can be used to address the issue of whether there is any suggestion that the assumptions underlying this model are not valid. *[2 Marks]*

(viii) What do you conclude? *[2 Marks]*

Minitab Output for Question 2.

```
The regression equation is
mpg = 9.67 + 0.276 mph

Predictor        Coef      SE Coef          T        P
Constant       9.6656       0.5957      16.23    0.000
mph           0.27621      0.01987      13.90    0.000

S = 0.8922       R-Sq = 93.7%      R-Sq(adj) = 93.2%

Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1      153.75      153.75     193.16    0.000
Residual Error    13       10.35        0.80
Total             14      164.10

Unusual Observations
Obs        mph         mpg         Fit      SE Fit     Residual     St Resid
 12       35.7      21.800      19.526       0.281        2.274        2.68R
```





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?

[2 Marks]

b)

An auctioneer's office have compiled data on the area X_1 (in sq. feet), the age X_2 (in years) and the price Y (in €) of a random sample of houses from a particular region. A matrix scatterplot of the data is provided below.

(i) What does the matrix scatterplot suggest to you?

[2 Marks]

(ii) The following normal errors regression model (Model 1)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

was fitted to the data. Interpret in your own words β_1 , β_2. Using the relevant output below make a decision as to whether age and area are useful predictors of price by giving the results of the significance tests (i.e. state the null and alternative hypothesis, test statistic, degrees of freedom, P-value, and conclusion) reported in the output.

[8 Marks]

(iii) A second model (Model 2)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

was fitted which included the variable X_3 representing the interaction between the age and area of the house.

Interpret β_3 in your own words.

[2 Marks]

(iv) Based on the output included for Model 2 below, what can you conclude at this stage of the analysis as to whether age should be dropped from the model?

[2 Marks]

(v) An additional variable "Garage Space" X_4 was collected in order to investigate the influence the presence of a shed or garage has on the price. The data for this variable were coded as

$$X_3 = \begin{cases} 1\text{- when there was neither a shed nor garage present} \\ 2\text{ - when a shed alone was present} \\ 3\text{ - when a garage was present} \end{cases}$$

A third model (Model 3) was fitted which included the new variable Garage Space

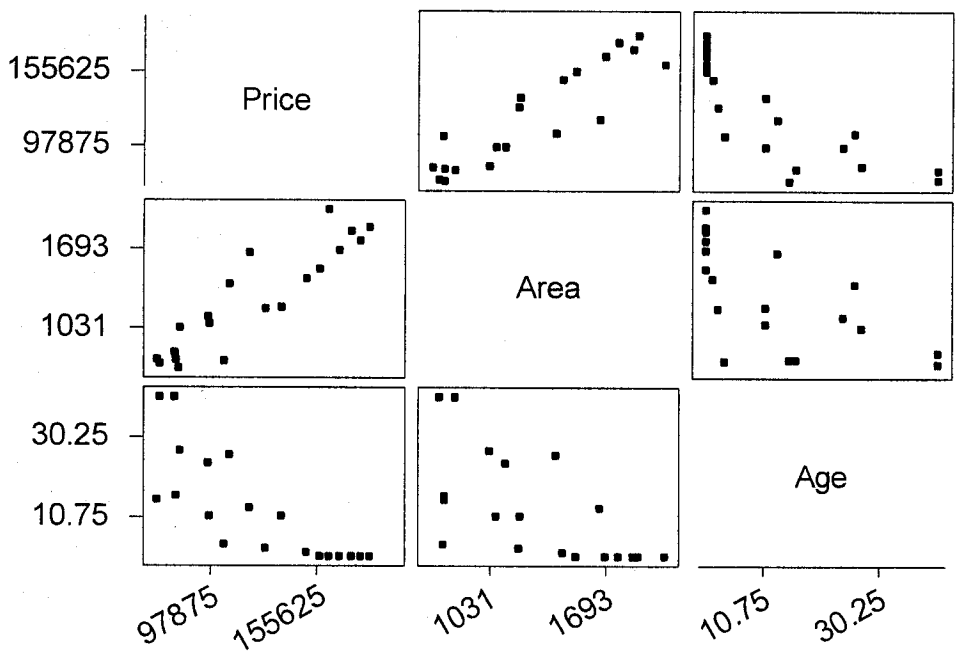
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

and the output is included below.

Interpret β_4 in your own words and indicate any problems you might have with including this variable in the model in its current form?

[4 Marks]

Minitab Output for Question 3.



Model 1.

The regression equation is
Price = 57544 + 61.5 Area - 1146 Age

19 cases used 1 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	57544	15970	3.60	0.002
Area	61.509	9.860	6.24	0.000
Age	-1146.0	318.9	-3.59	0.002

S = 13721 R-Sq = 89.5% R-Sq(adj) = 88.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	25583291718	12791645859	67.94	0.000
Residual Error	16	3012336786	188271049		
Total	18	28595628503			

Model 2.

The regression equation is
Price = 47196 + 69.9 Area + 314 Age - 1.46 Area*Age

19 cases used 1 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	47196	15817	2.98	0.009
Area	69.85	10.17	6.87	0.000
Age	314.1	829.9	0.38	0.710
Area*Age	-1.4596	0.7750	-1.88	0.079

S = 12744 R-Sq = 91.5% R-Sq(adj) = 89.8%

Model 3.

The regression equation is
Price = 38580 + 55.0 Area + 521 Age - 1.50 Area*Age + 12590 Garage Space

19 cases used 1 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	38580	16808	2.30	0.038
Area	55.00	15.11	3.64	0.003
Age	521.4	826.5	0.63	0.538
Area*Age	-1.5013	0.7581	-1.98	0.068
Garage S	12590	9644	1.31	0.213

S = 12455 R-Sq = 92.4% R-Sq(adj) = 90.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	26423792067	6605948017	42.58	0.000
Residual Error	14	2171836436	155131174		
Total	18	28595628503			

Question 4.

(i)

(a) Define multicollinearity and give three ways to identify the presence of multicollinearity in a multiple regression model.

[3 Marks]

(b) What problems can be caused by multicollinearity?

[2 Marks]

(ii)

A sales manager for a large corporation wishes to evaluate the performance of the company's sales representatives. Each sales representative is solely responsible for one sales territory, and the manager decides that it is reasonable to measure the performance, Y, of a sales representative by using the yearly aggregate sales (in units) of the corporation's products in the representative's sales territory. The manager feels that sales performance Y substantially depends on eight independent variables:

X_1 = the number of months the representative has been employed by the company.

X_2 = unit sales of the company's product and competing products in the sales territory.

X_3 = advertising expenditure in the territory.

X_4 = weighted average of the company's market share in the territory for the previous years.

X_5 = change in the company's market share in the territory over the previous four years.

X_6 = number of accounts handled by the representative.

X_7 = average workload per account.

X_8 = sales manager's ratings of the representative.

(a) What is the sample size?

[1 Mark]

(b) Use the matrix scatterplot and correlation matrix to carry out a preliminary assessment of the relationship between Sales and each of the 8 predictors.

[2 Marks]

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

[2 Marks]

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

[4 Marks]

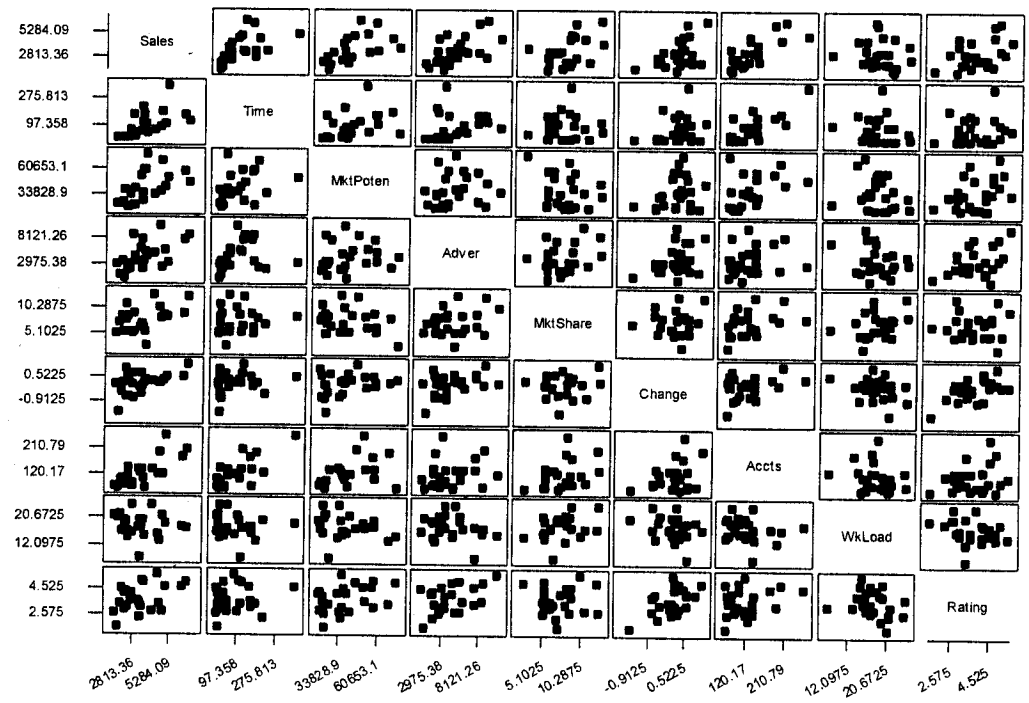
(e) What additional information would you need to complete this analysis?

[2 Marks]

(f) Summarise your findings for the sales manager.

[4 Marks]

Minitab Output for Question 4.



	Sales	Time	MktPoten	Adver	MktShare	Change	Accts	WkLoad
Time	0.623							
	0.001							
MktPoten	0.598	0.454						
	0.002	0.023						
Adver	0.596	0.249	0.174					
	0.002	0.230	0.405					
MktShare	0.484	0.106	-0.211	0.264				
	0.014	0.613	0.312	0.201				
Change	0.489	0.251	0.268	0.377	0.085			
	0.013	0.225	0.195	0.064	0.685			
Accts	0.754	0.758	0.479	0.200	0.403	0.327		
	0.000	0.000	0.016	0.338	0.046	0.110		
WkLoad	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
	0.577	0.391	0.212	0.188	0.087	0.163	0.341	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277
	0.046	0.631	0.078	0.041	0.911	0.004	0.272	0.180

Cell Contents: Pearson correlation
P-Value