MATH38141 Regression Analysis – Coursework

This coursework accounts for 20% of overall mark for this course and it may take around 10 hours to complete. Please present your solution in the form of a report, which you should upload on Blackboard as a single file before the deadline. You can use R to perform your calculations, but you must show the formulae in the text (not as R code) that you have used for the calculations. Marks will be awarded for correct and accurate calculations and their interpretation. Interpretations should be explained in words, referring to the context of the exercise, rather than naming generic symbols only. High marks will be less likely if the presentation of the results is unclear, too short or unnecessarily long and confusing, or if any formulas used in the calculations are missing from the text.

Submit your solution as a single file to Blackboard by 6pm on Wednesday, November 22, 2023.

1. Jane, an amateur [redacted] on the success of her top 10 favourite [redacted] ving 4 variable for each film:

   - *BoxOffice* – [redacted] es) in the first year, in million of US
   - *Production* – [redacted]
   - *Promotion* – [redacted]
   - *Books* – Tota[redacted] ooks the movie is based on), in millio[redacted]

   A tab-delimited t[redacted], is available on Blackboard.

   (a) Draw scatte[redacted] variables. Describe any observable t[redacted]

   (b) Formulate a[redacted] using *BoxOffice* as the response and[redacted]

   (c) Calculate the LSEs and construct 95% confidence intervals for all regression coefficients.

   (d) Provide an interpretation for the estimated coefficients obtained in (c).

   (e) Calculate and provide an interpretation for the R2 statistic for the model.

   (f) Jane argues that, when fitting a multiple linear regression model to the data using *BoxOffice* as the response and the other variables as the explanatory variables, the intercept term β_0 should be set to zero. Is this argument reasonable? Why?

   (7 marks)

   Excited about discovering more about her favourite film, Jane decides to test a theory and see whether the success of the film is really linked to the success of the book, or whether one might just need to know about the amount of money invested in producing and advertising the film. To investigate whether *Books* also affects *BoxOffice*, Jane fits two multiple linear regression models to the *BoxOffice* data:

- Model 1, with explanatory variables *Production* and *Promotion*;
- Model 2, with explanatory variables *Production*, *Promotion* and *Books*.

(g) Decide which one is the reduced model. Then fill in the following ANOVA table to compare the nested models.

Source	s.s.	d.f.	m.s.	F-ratio
Regression fitting reduced model	?	?	-	-
Extra	?	?	?	?
Residual fitting full model	?	?	?	-
Total	?	?	-	-

(h) Calculate the p-value associated with the significance of *Books*. Do you think *Books* should be included in the multiple linear regression model?

(i) Regressing *B* e significance of *Books* under this si n contradict that given in (h)? Com

(4 marks)

2. A dataset concer e USA. It contains the following variable

- ANS: Annua
- NSF: Numb
- INV: Invent (in thousands of $);
- ASA: Amou
- SSD: Size of
- NCS: Numb

A tab-delimited t , is available on Blackboard.

A multiple linear regression model Ω is proposed to describe the relationship between the response variable ANS and the other 5 explanatory variables (NSF, INV, ASA, SSD, NCS).

A retail expert believes, however, that the variation in ANS can be adequately explained by the variable INV alone, and hence proposes a simple linear regression model ω for the data.

(a) Specify the models Ω and ω, and state the model assumptions clearly.

(b) Calculate the residual sums of squares fitting Ω and ω respectively.

(c) Explain why in (b) the residual sum of square of Ω is not larger than that of ω.

(d) Under model Ω, test whether the regression coefficients of ASA and SSD are 15 and 10, respectively, at the 10% significance level, and explain your conclusions.

(e) Suppose that we want to compare how well two new shops in two locations will perform:

shop	NSF	INV	ASA	SSD	NCS
1	3.0	500	5.0	5.0	5
2	5.0	500	10.0	10.0	10

Calculate the predicted difference in annual net sales between shop 2 and shop 1. Do we predict the two shops to perform significantly differently at a 5% significance level?

(f) It is suggested that the relationship between ANS and INV depends on the number of competing stores in the district, i.e. on NCS.

i. Propose a new model Ω_1 to reflect this suggestion, making sure model ω is nested within Ω_1. Exclude the other regressor variables (i.e. NSF, ASA and SSD).

ii. Carry out a hypothesis test to compare ω against Ω_1 and make conclusions.

iii. Based on the fitted model Ω_1, plot four fitted regression lines on the same diagram to display the relationships between ANS and INV for the four values of NCS of 0, 4, 8 and 12, respectively.

Comme  and INV for these four different

(9 marks)

[**Total: 20 marks**]