STAT9004

Continuous Assessment 2

**Please hand the assignment in to me on (or before) – see Canvas notifications**

**This assignment is worth 30% of your final grade.**

**This analysis must be carried out using R.**

**Please hand in a written report as well as a copy of the R script used to generate the output referred to in the written report. You may also combine the report with the code using an R markdown document.**

Question 1

A researcher would like to explore the relationship between the diameter of a tree trunk and the volume of the trunk for a particular species of tree. The tree.xlsx dataset is available on Canvas. The dataset consists of *n* observations of 2 variables, Diam and Vol which measure the diameter (m) and the volume (m3) of *n* tree trunks from the same species.

The aim of the question is to find out whether the variable Diam, is of use in predicting the value of Vol. Please write up the analysis in the form of a report.

1. Make a numerical and graphical summary of the data, commenting on the results. Include: boxplots, histograms, scatterplot and the correlation coefficient.
2. Fit a model of the form and interpret the value of . Note that you will need to consider the results from your exploratory data analysis in part (a) to fit a valid model.
3. Calculate a 95% confidence interval for the coefficient.
4. Test the hypothesis:

H0:

HA:

What do the results of the hypothesis test imply for the regression model?

1. Plot the regression line onto a scatterplot of the data and plot a 95% prediction band.
2. Plot the studentized residuals against the fitted values and identify any outliers.
3. Plot the leverage of each case and identify any observations that have high leverage.
4. Identify the observation that has the largest influence on the estimate of the coefficient. Explain why this observation has a large influence.

[30 marks]

Question 2

The div dataset is available on Blackboard. The dataset consists of 77 observations and 7 variables. The first 4 observations of the dataset are shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| year | divorce | unemployed | femlab | marriage | birth | military |
| 1920 | 8 | 5.2 | 22.70 | 92.0 | 117.9 | 3.2247 |
| 1921 | 7.2 | 11.7 | 22.79 | 83.0 | 119.8 | 3.5614 |
| 1922 | 6.6 | 6.7 | 22.88 | 79.7 | 118.2 | 2.4553 |
| 1923 | 7.1 | 2.4 | 22.97 | 85.2 | 110.5 | 2.2065 |

Variables

year the year from 1920-1996

divorce per 1000 women aged 15 or more

unemployed unemployment rate

femlab percent female participation in labour force aged 16+

marriage marriages per 1000 unmarried women aged 16+

birth births per 1000 women aged 15-44

military - military personnel per 1000 population

Please answer the questions below and write up the analysis in the form of a report.

1. Make a numerical and graphical summary of the data, commenting on the results. Include: boxplots, histograms, scatterplots and correlation coefficients.
2. Fit the model :
3. Interpret the coefficient for femlab.
4. Calculate the variance inflation factors for this model and discuss their implications for collinearity in the model.
5. By fitting alternative models, determine whether collinearity has an impact on the coefficient estimates.
6. Create a partial regression plot to examine relationship between birth and divorce adjusted for unemployed, femlab, marriage and military.
7. Test the hypothesis:

H0:

HA: at least one of the

What do the results of the hypothesis test imply for the regression model?

1. Assess the fit of the model using diagnostic plots, commenting on the assumptions of the regression model and influential points.
2. Use an F-test to compare the full model to the model including all variables except unemployment.
3. Compare the predictive accuracy of the two models from part (c) using 50 repeats of 10-fold cross validation.
4. Can you suggest any improvements to the model?

[35 marks]

Question 3

1. Please write two or three paragraphs on the concept of step-wise regression, including a brief description of three different types of step-wise regression and an overview of the different criteria that can be used to decide whether an explanatory variable is to be retained in the final model. Finally, select and discuss one problem associated with step-wise regression.
2. Select a step-wise regression function from the statistical software R and give a brief description on how the algorithm is implemented. Include a description of the type of step-wise regression and the criterion used to determine whether a variable is included in the final model.
3. In class we ran simulations to explore the performance of stepwise regression for model selection. We investigated how the number of predictors that are linearly related to the outcome affected the number of variables that were incorrectly retained in the model or incorrectly omitted from the model. We also examined how the strength of the relationship between the predictors and the response affected the number of variables that were incorrectly retained in the model or incorrectly omitted from the model. Can you expand this investigation further? You can choose from the options below or explore your own idea. If you would like to investigate an idea that is not listed then please let me know before commencing the analysis. Present the findings clearly using tables or graphs and discuss the implications of your findings.

Choose one from the following:

* How does varying the number of predictors affect the performance of step-wise regression?
* How does varying the size of the data set relative to the number of predictors affect the performance of step-wise regression?
* How do different step-wise regression algorithms compare?
* How does collinearity of the predictors impact the performance of step-wise regression?
* Suppose the aim of the model is statistical inference and we are interested in the values and standard errors of the coefficients. How does the final model (selected using step-wise regression) compare to simply using the full model?
* Use 10-fold cross validation with step-wise regression to select a final model based on minimising the RMSE. Does using cross-validation prevent spurious conclusions?
* Compare the performance of step-wise regression with LASSO regression.

[35 marks]