MACQUARIE University
SYDNEY·AUSTRALIA

STAT2170 AND STAT6180
APPLIED STATISTICS

Part I Practice Assignment S2 2023

Question 1 [45 marks]

World Health Organisation's (WHO) specialised cancer agency, the International Agency for Research on Cancer (IARC) has designated fine particulate matter ($PM_{2.5}$) as carcinogenic to human beings. $PM_{2.5}$ particles have a diameter of 2.5 micrometers (0.0025 mm) or smaller and they are small enough for people to breath them deeply into lungs and sometimes $PM_{2.5}$ particles can even enter the bloodstream. Research indicates that temperature in degrees ($^{\circ}C$), relative humidity in percentage (%), wind speed in kilometers per hour (km/h), and precipitation in millimeters (mm) are potential predictors for $PM_{2.5}$ concentration in milligram per cubic meter ($\mu g/m^3$).

A random sample of the annual mean temperature, humidity, wind, precipitation and $PM_{2.5}$ concentration at 56 test locations was collected. The data is available in the file `pm25.csv` on iLearn.

Variable	Description
temperature	The annual mean temperature in degrees
humidity	The annual mean relative humidity in percentage
wind	The annual mean wind speed in kilometers per hour
precipitation	The annual mean precipitation in millimeters
pm25	The annual mean $PM_{2.5}$ concentration in milligram per cubic meter

a. [**7 marks**] Produce a plot and a correlation matrix of the data. Comment on possible relationships between the response and predictors and relationships between the predictors themselves.

b. [**6 marks**]

- Fit a model using all the predictors to explain the `pm25` response.
- Using the full model, estimate the impact of humidity on $PM_{2.5}$ concentration. Do this by producing a 95% confidence interval that quantifies the change in $PM_{2.5}$ concentration for each extra percentage of relative humidity and comment.

c. [**14 marks**] Conduct an F-test for the overall regression i.e. is there **any** relationship between the response and the predictors. In your answer:

- Write down the mathematical multiple regression model for this situation, defining all appropriate parameters.
- Write down the Hypotheses for the Overall ANOVA test of multiple regression.
- Produce an ANOVA table for the overall multiple regression model (One combined regression SS source is sufficient).
- Compute the F statistic for this test.
- State the Null distribution for the test statistic.
- Compute the P-Value
- State your conclusion (both statistical conclusion and contextual conclusion).

d. [**10 marks**] Validate the **full** model and comment on whether the full regression model is appropriate to explain the $PM_{2.5}$ concentration at various test locations.

e. [**2 marks**] Find the R^2 and comment on what it means in the context of this dataset.

f. [**3 marks**] Using model selection procedures discussed in the course, find the best multiple regression model that explains the data. State the final fitted regression model.

g. [**3 marks**] Comment on the R^2 and adjusted R^2 in the full and final model you chose in part f. In particular explain why those goodness of fitness measures change but not in the same way.

Question 2 [25 marks]

A business wants to advertise their product in Film media by using product placement in a movie. To maximise the brand recognition from the placement, the business conducted a study recording the correct number of brands identified by individuals in an experiment that watched different types of movies. Each movie in this experiment featured six different brands.

Variable	Description
Gender	Gender of the individual watching the movie
Genre	Genre of the movie being watched
Score	The number of correct brands recalled by the individuals after the movie

The data is available in the file `movie.csv` on iLearn.

a. [**2 marks**] For this study, is the design balanced or unbalanced? Explain why.
b. [**8 marks**] Construct two different preliminary graphs that investigate different features of the data and comment.
c. [**4 marks**] Write down the **full** mathematical model for this situation, defining all appropriate parameters.
d. [**9 marks**] Analyse the data to study the effect of `Gender` and `Genre` on the brand recall Score. These conclusions are only required to be at the qualitative level and can be based off the outcomes of the hypothesis tests you conducted in this part and the preliminary plots in part b. You do not need to statistically examine the multiple comparisons between contrasts and interactions. Remember to
    - state the null and alternative hypothesis for each test, and
    - check assumptions.
e. [**2 marks**] Based on your results from part d), discuss the practical implications of your findings for the business that aims to maximise the brand recognition from the placement. What advice/interpretation would you provide on the effect drama genre on the brand recall Score.