STAT2401 Analysis of Experiments
Semester 1, 2025
Assignment 2 (15%, 50 marks)

Due date: Week 8 Friday 16th of May 2025 by 11:59pm

- Working in **pairs** of two students is strongly encouraged. If you work as a pair:
  - submit only one assignment per team;
  - each student must contribute towards ALL questions;
  - list the team members by name, student ID and **state what each team member has contributed to the assignment before answering the questions. Your assignment will not be marked without this statement.**
- You must submit **two files (the main document and Rmarkdown)** using **two separate LMS submission button**:
  1. The main document consists of your answers is to be submitted to the first **Turnitin LMS** button as a **pdf** file only. Unlimited versions can be submitted until the due date and must comply to the following requirements:
     - You are only required to submit your submission **in PDF** with the following **approximate layout and order**:

Page 1-6	Question 1
Page 7-12	Question 2

     - Use font size **12** for answer or comment.
     - Use smaller font size of (eg 9) for R-code snippets, figures or `R`-outputs.
     - Your submission main document should have no more than **12** pages.
     - Do not include any cover sheet, cover page or title page since that should be counted as 1 page.
     - Resize the plots accordingly to fit the space or combine the plots whenever applicable, using par(mfrow) command.
     - **Your submission will not be marked if you don't follow the layout and order**.
     - Only electronic submission through LMS is acceptable.
  2. Rmarkdown (no need to be knitted) attachment that contains the complete R code as your working is to be submitted into the second LMS button within File Response.
- **Marking of late assignments will follow the university rules.**

Assignment Questions

The data required for this assignment can be found within Assignment 2 folder on the LMS.

1. **Air Pollution and Mortality. Does pollution kill people? (30 marks)**

   The dataset *Pollution.csv* obtained from one early study that explored this issue of pollution came from 60 Standard Metropolitan Areas in the US over the years 1959-1961. Total age-adjusted mortality from all causes, in deaths per 100,000 population (*Mortality*), is the response variable.

   Researchers collected four climate-related variables, eight socioeconomic variables, and three air-pollution variables. Their primary question concerned the effects, if any, of air pollution on mortality, after accounting for climate and socioeconomic differences among the cities.

   The fifteen explanatory variables listed as follows:

   - *Over65:* percentage of the population aged 65 years or over;
   - *House:* population per household;
   - *Educ:* median number of school years completed by persons of age 25 years or more;
   - *Sound:* percentage of the housing that is sound with all facilities;
   - *Density:* population density (in persons per square mile of urbanized area);
   - *NonWhite:* percentage of 1960 population that is nonwhite;
   - *WhiteCol:* percentage of employment in white-collar occupations;
   - *Poor:* percentage of households with annual income under $3,000 in 1960;
   - *Precip:* mean annual precipitation (in inches);
   - *Humidity:* percent relative humidity (annual average at 1 P.M.);
   - *JanTemp:* mean January temperature (in degrees Fahrenheit);
   - *JulyTemp:* mean July temperature (in degrees Fahrenheit);
   - *HC:* relative pollution potential of hydrocarbons (HC);
   - *NOX:* relative pollution potential of oxides of nitrogen (NOX); and
   - *SO2:* relative pollution potential of sulphur dioxide (SO2).

   The explanatory variables can be grouped as measures of pollution, climate measurements and socioeconomic variables as follows:

   - *Climate*: Precip, Humidity, JanTemp and JulyTemp;
   - *Socioeconomic*: Over65, House, Educ, Sound, Density, NonWhite, WhiteCol and Poor;
   - *Pollution*: HC, NOX and SO2.

   It is desired to determine whether the pollution variables are associated with mortality after the other climate and socioeconomic variables are accounted for.

   To address the research question of interest you need to first select a good-fitting regression model.

(a) (5 marks) Carry out exploratory data analysis (EDA) of this dataset, taking into account the following groups of associations with the response:

- 1 plot for *Climate*
- 1 plot for *Socioeconomic*
- 1 plot for *Pollution*
- 1 plot combining the comparisons of the response against *State.Code* and *Region* respectively.

Hint. Your answer must include 4 plots for EDA including interpretation in a few concise sentences. Do not include the R code.

(b) (9 marks) Perform model selection process using *All Subset* method to arrive at good regression models that account for variation in mortality between the cities that can be attributed to differences in *climate and socioeconomic factors.* Identify optimal model(s) based on each of the adjusted R^2, BIC and C_p.

Hint. Use the subset size of 10 variables for the best 1 criteria. The optimal models must state the number and names of the selected variables.

Hint. Your answer must include 1 plot for model selection, R output from outmat and their interpretations. Do not include the R code.

(c) (4 marks) Fit an optimal model with the lowest C_p as obtained in (b). Write the equation of this fitted model. Interpret this model.

Hint. Do not include the R code.

(d) (7 marks) Perform diagnostics checking on the model in (c). Do you think there are influential points in the data? Identify the cities which are influential points using leverage and Cook's distance respectively.

Hint. Use 'plot(file.lm)' and 'par(mfrow=c(2,2))' for diagnostics plots.

Hint. Your answer must include at most 3 plots, results and comment. Use the interval of (-2,2) for standardised residuals.

(e) (5 marks) Using the model obtained in (c), add the three pollution variables (transformed to their natural logarithm) and obtain the p-value from the extra-sum-of-squares F-test due to their addition. Summarise your findings in a few concise sentences.

Hint. Your answer must include a snippet of R code for adding variables.

2. **Body Measurements (25 marks)**

The dataset *body.csv* contains the response variable 'Weight', 23 quantitative predictors that are girth measurements of different body parts along with age and height, and a single categorical variable ('Gender': 1 for Male, 0 for Female).

Hint. Use the function 'names' to list the variable names. This is a data frame with 507 observations on 25 variables, with the detail can be found in the following link https://www.openintro.org/data/index.php?data=bdims

(a) (4 marks) Carry out exploratory data analysis (EDA) of this dataset before you do any modelling.

Hint. Your answer must include at most 2 plots for EDA including interpretation in a few concise sentences. Do not include the R code.

(b) (10 marks) After the exploratory analysis has been carried out, split the dataset into a training set and a testing set so that the training set contains 80% of the data and the testing set contains 20%. Construct a multiple linear regression model for this dataset using the training set to create 2 final fitted models at a significance level of 10%, based on the following variable selection methods :

- (Model 1) The Forward selection;
- (Model 2) The Backward selection.

Write the two fitted model equations and compare them in a few concise sentences.

Hint. You must use set.seed(2401) for reproducibility. Your answer must include a snippet of R code for the splitting.

Hint. Use the R command step() based on the AIC criterion for the Forward and Backward methods.

Hint. Use a significance level of 10% for the final fitted models. Your answer must include the R output model summary, fitted model equations and comment.

(c) (5 marks) Perform diagnostics checking for each of the final fitted models, Model 1 and Model 2 respectively.

Hint. Use 'plot(file.lm)' and 'par(mfrow=c(2,2))' for diagnostics plots. Your answer must include the plots and comment. Do not include the R code.

Hint. Use the interval of (-2,2) for standardised residuals whenever applicable.

(d) (6 marks). Despite any inadequacies that you may or may not have identified above, you use the two models obtained in (b) to make predictions of *Weight* in the test set.

(i) Produce a correctly drawn and labelled plot of predicted values against the actual values in the test set, and obtain the root mean squared error of prediction (RMSEP) based on each fitted model.

(ii) Using the RMSEPs and the plots you produced, comment on how well the models performed.

Hint. Your answer must include a snippet of R code to calculate and plot the RMSEPs, the plots and comment.