assignment_02

October 2, 2025

1 Assignment 2

---

1. **Multivariate Data Analysis & Vizualisation** *(11)*

For the `vehicles` dataset from the `fueleconomy` package in `R`,

a. develop the following plots *(10)*

- histogram for highway & city MpG
- box plot for highway & city MpG
- bar plot for mean highway & city MpG vs. vehicle year
- line plot for mean highway & city MpG vs. engine cylinders
- scatter plot for highway & city MpG vs. engine displacement

b. evaluate the correlation between *(1)*

- engine displacement and highway & city MpG
- engine cylinders and highway & city MpG
- vehicle year and highway & city MpG

[ ]:
```r
# Vehicles Dataset - Fuel Economy Package
## Load Packages
library(dplyr)
library(ggplot2)
library(patchwork)
library(fueleconomy)

# Load the dataset (filtered for non-zero & non-NA values; augmented with new variable 'I' indicating if vehicle is manufactured after 2009)
data <- fueleconomy::vehicles %>%
  filter(year != 0, cyl != 0, displ != 0, is.na(year) == FALSE, is.na(cyl) == FALSE, is.na(displ) == FALSE) %>%
  mutate(I = ifelse(year > 2009, 1, 0))
```

---

2. **Linear Regression** *(18)*

For the `vehicles` dataset from the `fueleconomy` package in `R`,

a. develop the following models *(5)*

- $\texttt{cty} = \beta_0 + \beta_1 \texttt{year} + \beta_2 \texttt{cyl} + \beta_3\ \texttt{displ}$
- $\texttt{hwy} = \beta_0 + \beta_1 \texttt{year} + \beta_2 \texttt{cyl} + \beta_3\ \texttt{displ}$

b. for the two models, compute *(6)*

- Sum of Squares Total (SST)
- Sum of Squares Regression (SSR)
- Sum of Squared Errors (SSE)
- Residual Standard Error (RSE)
- R-squared (R^2)
- Adjusted R-squared ($\bar{R}^2$)

c. for the two models, *(7)*

- develop residuals plot
- comment upon the validity of the assumptions of linear regression

(Hint: To comment upon multicollinearity, develop pairwise correlation for the exogneous variables)

---

3. **Symbolic Regression** *(5)*

For the `vehicles` dataset from the `fueleconomy` package in `R`, compare the following models explored via symbolic regression with the linear regression model developed in the previous question *(5)*

- $\texttt{cty} = \beta_0 + \beta_1\ \text{I}(\texttt{year} \geq 2010) + \beta_2\ \log(\texttt{displ})$
- $\texttt{hwy} = \beta_0 + \beta_1\ \text{I}(\texttt{year} \geq 2010) + \beta_2\ \log(\texttt{displ})$

(Hint: Compare the two set of models using model statistics and residual plots)

---

4. **Logistic Regression** *(11)*

For the `TravelMode` dataset from the `AER.` package in `R`, explore how alternate-specific variables (`travel`: in-vehicle travel time, `wait`: terminal waiting time, `vcost`: vehicle operational cost, and `gcost`: generalized travel cost) as well as individual-specific variables (`income`: household income, and `size`: traveling party size) impact choice of travel mode (`air`, `train`, `bus`, `car`). To this end,

(Note: The data is availble in **long format** with one row per *individual–mode* combination.)

a. develop the model *(4)*

b. compute the following statistics *(7)*

- log-likelihood for the
  - equally likely model

    - market share model
    - estimated model
- estimated model R-squared with respect to the
    - equally-likely model
    - market share model
- estimated model adjusted R-squared with respect to the
    - equally-likely model
    - market share model

[ ]:
```r
# TravelMode Dataset - AER Package
## Load Packages
library(AER)
library(dplyr)
library(mlogit)

## Load the dataset (choices mutated from yes/no to TRUE/FALSE)
data("TravelMode", package = "AER")
long_data <- TravelMode %>%
  mutate(choice = choice == "yes")

model_data <- mlogit.data(
  long_data,
  choice = "choice",
  shape = "long",
  chid.var = "individual",
  alt.var = "mode"
)
```