STAT5002 Lab13 Question Sheet

Introduction to Statistics

STAT5002

1 `mtcars` data

The `mtcars` dataset is a built-in dataset in R containing specifications and performance data for 32 car models from the 1970s. Below is a description of the variables selected for analysis:

Variable	Description
mpg	Miles per gallon — a measure of fuel efficiency (response variable).
cyl	Number of cylinders in the engine (typically 4, 6, or 8).
disp	Engine displacement (in cubic inches) — a measure of engine size.
hp	Gross horsepower — a measure of engine power.
wt	Vehicle weight in 1000 lbs — heavier cars tend to consume more fuel.
qsec	1/4 mile time (in seconds) — time taken to travel a quarter mile from a standstill (acceleration performance).

The following R code builds a new dataframe using the selected variables.

```r
selected_vars <- c("mpg", "cyl", "disp", "hp", "wt", "qsec")
dat <- mtcars[, selected_vars]
str(dat)
```

```
'data.frame':   32 obs. of  6 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
```

Using the new dataframe `dat`, we want to build a multiple linear regression model to predict `qsec` (1/4 mile time) using other variables as independent variables. We will apply both backward and forward variable selection methods, using F-tests at a 5% significance level to guide the model selection.

1.1 Plot the pairwise scatter plots and calculate the pairwise correlation coefficients.

- Do we need to be concerned about the effect of multicollinearity in this dataset?

1.2 Select the best model using the backward selection.

- We start the backward selection the full model containing all variables.

1.3 Select the best model using the forward selection.

- We start the forward selection with a baseline model only containing the intercept.

- Is the selected model the same as the one obtained through backward selection?

1.4 Write down the fitted model,

- How should the fitted model be interpreted?
- Does the fitted model align with intuition?

2 F-test

Only attempt this question if time permits during the lab session; otherwise, please prioritise completing Question 3 first. This question is intended for practicing the F-test and will not be included in the final assessment.

Compare a reduced model that includes only the explanatory variables `cyl`, `wt` and `hp` with the full model (which includes all available predictors) using the F-test. Are the additional variables in the full model significant in explaining the dependent variable (`qsec`) at the 5% level of significance?

2.1 Specify the hypotheses in words

Hint: determine first the null model and the alternative model.

2.2 Calculate the observed F-statistic

Recall the F-statistic:

$$F = \frac{(\widehat{SSE}_{H_0} - \widehat{SSE}_{H_1})/(p-q)}{\widehat{SSE}_{H_1}/\,(n-(p+1))} \sim F_{p-q,n-(p+1)}.$$

Hint: what are the values of p, q, and n?

2.3 Calculate the P-value and draw conclusion

- You can use the function `pf()` for the P-value.

3 Logistic regression

A local health clinic sent out fliers to its clients to encourage everyone – especially older individuals at high risk of complications – to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot.

Additional data were collected on each client's age (x_1) and health awareness. The latter was summarized into a health awareness index (x_2), where higher values indicate greater awareness.

Clients who received a flu shot were coded as $y_i = 1$, and those who did not were coded as $y_i = 0$. The data were imported into R below.

```
dat = read.csv("data/flushots.csv", header = T)
```

3.1 Fit a logistic regression model

- Fit the model using R
- Write down the estimated regression equation.
- Interpret the regression coefficients associated with x_1 and x_2 in terms of the odds.

3.2 Model prediction

- What is the estimated odds and the estimated probability that clients aged 55 with a health awareness index of 60 will receive a flu shot?
- How do you interpret the estimated odds?

Hint: using `type="response"` gives the predicted probability, without specifying `type`, you will get the predict log-odds