STAT5002 Lab13 Solution 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 ...
```

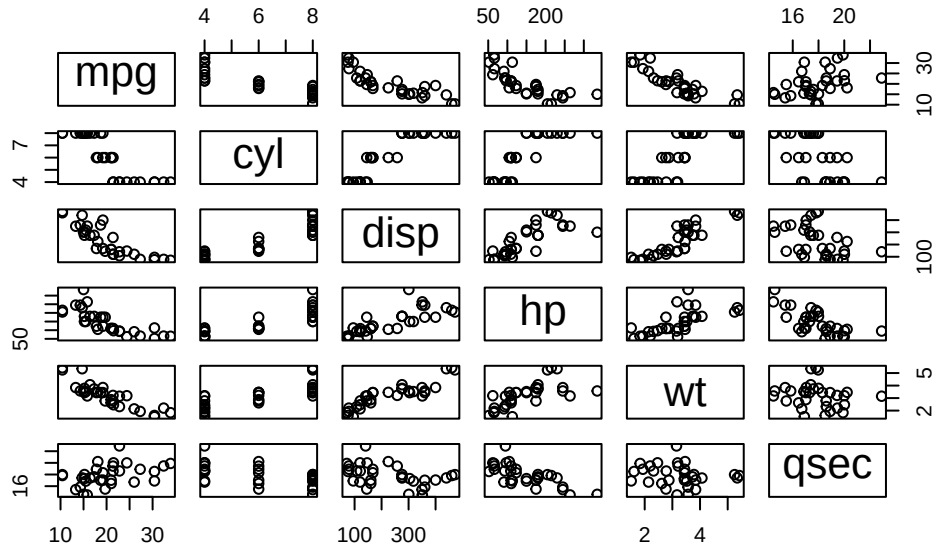
Using the new dataframe `dat`, we want to build a multiple linear regression model to predict `qsec` (1/4 mile time) using other variables as independent variables. We will apply both backward and forward variable selection methods, using F-tests at a 5% significance level to guide the model selection.

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?

Solution:

```
pairs(dat)  # Pairwise scatter plots
```




```
cor(dat)  # Pairwise Pearson correlation matrix
```

```
            mpg        cyl       disp         hp         wt       qsec
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684 -0.8676594  0.4186840
cyl  -0.8521620  1.0000000  0.9020329  0.8324475  0.7824958 -0.5912421
disp -0.8475514  0.9020329  1.0000000  0.7909486  0.8879799 -0.4336979
hp   -0.7761684  0.8324475  0.7909486  1.0000000  0.6587479 -0.7082234
wt   -0.8676594  0.7824958  0.8879799  0.6587479  1.0000000 -0.1747159
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234 -0.1747159  1.0000000
```

We see high correlation among these pairs of independent variables, for example:

- `disp` and `hp` – larger engines (higher disp) often produce more power (hp)
- `wt` and `disp` – larger engines tend to be in heavier cars
- `cyl` and `disp` – more cylinders typically mean larger displacement
- `cyl` and `hp` – more cylinders typically mean more power as well
- `mpg` is also highly correlated to `disp`/`cyl`/`wt`

Overall, there are high correlation between almost all pairs of independent variables. This suggests multicollinearity, model selection should be applied.

1.2 Select the best model using the backward selection.

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

Solution:

- start from the full model

```
BM1 = lm(qsec ~ ., data = dat)
drop1(BM1, test = "F")
```

```
Single term deletions

Model:
qsec ~ mpg + cyl + disp + hp + wt
       Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>              27.022  6.5894
mpg     1    0.2818 27.304  4.9214  0.2711 0.606986
cyl     1    3.0562 30.078  8.0182  2.9406 0.098277 .
disp    1    0.1520 27.174  4.7690  0.1463 0.705209
hp      1   12.0924 39.115 16.4242 11.6350 0.002125 **
wt      1   10.4747 37.497 15.0727 10.0785 0.003836 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `disp` is the least significant independent variable, which should be dropped.

```
BM2 = update(BM1, . ~ . - disp)
drop1(BM2, test = "F")
```

```
Single term deletions

Model:
qsec ~ mpg + cyl + hp + wt
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              27.174  4.7690
mpg     1    0.2173 27.391  3.0238  0.2159 0.6459140
cyl     1    5.6835 32.858  8.8464  5.6471 0.0248368 *
hp      1   13.6552 40.829 15.7973 13.5676 0.0010162 **
wt      1   15.4960 42.670 17.2084 15.3967 0.0005411 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- mpg is now the least significant independent variable, which should be dropped.

```
BM3 = update(BM2, . ~ . - mpg)
drop1(BM3, test = "F")
```

```
Single term deletions

Model:
qsec ~ cyl + hp + wt
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              27.391  3.0238
cyl     1    7.0536 34.445  8.3561  7.2103 0.0120440 *
hp      1   15.8268 43.218 15.6169 16.1785 0.0003959 ***
wt      1   21.9453 49.337 19.8538 22.4328 5.702e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- All the P-values are less than 5%, so we stop, the model BM3 is the final model.

```
summary(BM3)
```

```
Call:
lm(formula = qsec ~ cyl + hp + wt, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.98267 -0.54986 -0.04903  0.52045  2.89743

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.768791   0.703685  28.093  < 2e-16 ***
cyl         -0.582570   0.216956  -2.685 0.012044 *
hp          -0.018812   0.004677  -4.022 0.000396 ***
wt           1.381334   0.291646   4.736  5.7e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9891 on 28 degrees of freedom
Multiple R-squared:  0.7233,    Adjusted R-squared:  0.6936
F-statistic:  24.4 on 3 and 28 DF,  p-value: 5.76e-08
```

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?

Solution:

- start from the baseline model

```r
FM1 = lm(qsec ~ 1, data = dat)
add1(FM1, scope = ~mpg + cyl + disp + hp + wt, test = "F")
```

```
Single term additions

Model:
qsec ~ 1
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              98.988 38.136
mpg     1    17.352 81.636 33.969  6.3767 0.0170820 *
cyl     1    34.603 64.385 26.373 16.1231 0.0003661 ***
disp    1    18.619 80.369 33.469  6.9501 0.0131440 *
hp      1    49.651 49.338 17.854 30.1902 5.766e-06 ***
wt      1     3.022 95.966 39.144  0.9446 0.3388683
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `hp` is the most significant independent variable, which should be added.

```
FM2 = update(FM1, . ~ . + hp)
add1(FM2, scope = ~mpg + cyl + disp + hp + wt, test = "F")
```

```
Single term additions

Model:
qsec ~ hp
       Df Sum of Sq    RSS     AIC F value   Pr(>F)
<none>              49.338 17.8544
mpg     1    4.2740 45.064 16.9549  2.7504 0.108007
cyl     1    0.0009 49.337 19.8538  0.0005 0.981671
disp    1    4.2289 45.109 16.9869  2.7187 0.109971
wt      1   14.8926 34.445  8.3561 12.5384 0.001369 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- in the new iteration, `wt` is the most significant independent variable, which should be added.

```
FM3 = update(FM2, . ~ . + wt)
add1(FM3, scope = ~mpg + cyl + disp + hp + wt, test = "F")
```

```
Single term additions

Model:
qsec ~ hp + wt
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>              34.445 8.3561
mpg     1    1.5873 32.858 8.8464  1.3527 0.25463
cyl     1    7.0536 27.391 3.0238  7.2103 0.01204 *
disp    1    2.8509 31.594 7.5915  2.5266 0.12317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- now `cyl` is the most significant independent variable, which should be added.

```
FM4 = update(FM3, . ~ . + cyl)
add1(FM4, scope = ~mpg + cyl + disp + hp + wt, test = "F")
```

```
Single term additions

Model:
qsec ~ hp + wt + cyl
       Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>              27.391 3.0238
mpg     1  0.217283 27.174 4.7690  0.2159 0.6459
disp    1  0.087549 27.304 4.9214  0.0866 0.7708
```

- All the P-values are greater than 5%, so we stop, the model `FM4` is the final model.

```
summary(FM4)$coefficients
```

```
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 19.76879067 0.703685081 28.093235 4.665901e-22
hp          -0.01881199 0.004676987 -4.022247 3.958639e-04
wt           1.38133373 0.291646252  4.736333 5.702169e-05
cyl         -0.58257004 0.216956429 -2.685194 1.204396e-02
```

```
summary(BM3)$coefficients
```

```
              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 19.76879067 0.703685081 28.093235 4.665901e-22
cyl         -0.58257004 0.216956429 -2.685194 1.204396e-02
hp          -0.01881199 0.004676987 -4.022247 3.958639e-04
wt           1.38133373 0.291646252  4.736333 5.702169e-05
```

- The models built by the forward and backward selection methods are identical.

1.4 Write down the fitted model,

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

Solution: Rounding to three decimal points, the fitted model is

$$\widehat{\text{qsec}} = 19.769 - 0.583 \cdot \text{cyl} - 0.019 \cdot \text{hp} + 1.381 \cdot \text{wt}$$

Holding other variables constant,

- every additional cylinder decreases the quater mile time by 0.583 seconds;
- an increase of one horsepower decreases the quater mile time by 0.019 seconds;
- an increase of one thousand pounds increases the quater mile time by 1.381 seconds.

This aligns with the intuition. More powerful cars are faster, heavier cars are slowers, while more number of cylinders somehow also reduces the quater mile time.

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.

Solution:

H_0 : The additional independent variables (`mpg` and `disp`) have no effect in explaining `qsec` after accounting for `wt` and `hp`.

H_1 : At least one of the additional independent variables (`mpg` and `disp`) has an effect in explaining `qsec`.

2.2 Calculate the observed F-statistic

Recall the F-statistic:

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

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

Solution: $p = 5$, $q = 3$, and n is the sample size

```r
alter.M = lm(qsec ~ ., data = dat)  # same as BM1
null.M = lm(qsec ~ hp + wt + cyl, data = dat)  # same as BM3
n = length(dat$qsec)
p = 5
q = 3
sse.h0 = sum(null.M$residuals^2)
sse.h1 = sum(alter.M$residuals^2)
nume = (sse.h0 - sse.h1)/(p - q)
deno = sse.h1/(n - (p + 1))
f.stat = nume/deno
f.stat
```

```
[1] 0.1776805
```

The P-value is greater than 0.05, which suggests that the data is consistent with H_0.

Note that you can use the `anova()` function to perform the partial F-test (the concept of ANOVA is not covered in this unit).

```r
anova(null.M, alter.M)
```

```
Analysis of Variance Table

Model 1: qsec ~ hp + wt + cyl
Model 2: qsec ~ mpg + cyl + disp + hp + wt
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 27.391
2     26 27.022  2   0.36933 0.1777 0.8382
```

2.3 Calculate the P-value and draw conclusion

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

Solution:

```r
pf(f.stat, p - q, n - (p + 1), lower.tail = F)
```

```
[1] 0.8382179
```

We fail to reject H_0 at the 5% significance level. That is, the additional variables in the full model do not significantly improve the model compared to the reduced model. The reduced model is adequate, and there is no strong evidence to justify including the extra variables.

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.

Solution:

```
model = glm(y ~ x1 + x2, data = dat, family = "binomial")
summary(model)
```

```
Call:
glm(formula = y ~ x1 + x2, family = "binomial", data = dat)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.58458    6.41824  -3.363 0.000771 ***
x1            0.22178    0.07436   2.983 0.002858 **
x2            0.20351    0.06273   3.244 0.001178 **
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 32.416  on 47  degrees of freedom
AIC: 38.416

Number of Fisher Scoring iterations: 6
```

The fitted model is

$$\widehat{\text{logit}(\text{p})} = -21.58458 + 0.22178 \times x_1 + 0.20351 \times x_2$$

- Holding other variable constant, the odds of a client who received a flu shot increase by about $\exp(0.22178) - 1 = 24.83\%$ with each additional year of age.
  - Or, holding other variable constant, an increase of 1 year in Age will increase the odds of a client who received a flu shot by a factor of $\exp(0.22178) \approx 1.2483$.
- Holding other variable constant, the odds of a client who received a flu shot increase by about $\exp(0.20351) - 1 = 22.57\%$ with each additional unit increase in health awareness index.

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

Solution:

```r
pred = data.frame(x1 = 55, x2 = 60)
log.odds = predict(model, pred)
log.odds
```

```
       1
2.823553
```

```r
odds = exp(log.odds)
odds
```

```
       1
16.83656
```

```r
prob = odds/(1 + odds)
prob
```

```
        1
0.9439354
```

```r
predict(model, pred, type = "response")
```

```
        1
0.9439354
```

A client aged 55 with a health awareness index of 60 is estimated to be 16.84 times more likely to receive a flu shot than not.