Name: Tan E-Zhen

Project: Individual R Programming Project for Course 1

The problem that I am investigating is how the displacement (cu.in.) (disp) in a car is affected by miles per gallon (mpg), number of cylinders (cyl), gross horsepower (hp), rear axle ratio (drat), weight (1000 lbs) (wt), ¼ mile time (qsec), type of engine (0 = V-shaped, 1 = straight) (vs), type of transmission (0 = automatic, 1 = manual) (am), number of forward gears (gear), and number of carburetors (carb). The dataset is obtained from the built-in dataset in R, mtcars.

The above variables will be assigned the following variables for simplicity:

Y = disp	X6 = qsec
X1 = mpg	X7 = vs
X2 = cyl	X8 = am
X3 = hp	X9 = gear
X4 = drat	X10 = carb
X5 = wt	

This problem will be investigated using linear regression method, using the question type based on the SUSS January 2018 Semester Examination Question 4[1]. The type of questions asked are as follows:

a) Describe the relationship between the dependent and independent variables using a linear equation.
b) Are the independent factors significant in influencing the dependent variable? Execute hypothesis tests at the 10% level of significance, to show which independent variable(s) is/are significant in influencing the dependent variable.

For part (a), to use R, the following code is used to create the linear model of disp as the dependent variable in relation to the other independent variables. A summary is then coded to view the results of the model. The results will tell us the coefficients of the independent variables, hence showing how the change in the independent variables will impact the dependent variable.

[1] Singapore University of Social Sciences (SUSS) (2018). BUS105e Examination – January Semester 2018.

```r
#describe the relationship between the independent and dependent variables
model.full=lm(disp ~ .,data=mtcars)
(summary.full=summary(model.full))
```

This code gives the following result:

```
Call:
lm(formula = disp ~ ., data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-72.28 -17.11  -0.23  18.95  55.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.812    228.061  -0.025  0.97991
mpg            1.940      2.598   0.747  0.46349
cyl           15.389     12.152   1.266  0.21924
hp             0.665      0.226   2.942  0.00778 **
drat           8.812     19.739   0.446  0.65987
wt            86.711     16.113   5.382 2.45e-05 ***
qsec         -12.974      8.623  -1.505  0.14730
vs           -12.115     25.258  -0.480  0.63643
am            -7.914     25.618  -0.309  0.76044
gear           5.127     18.058   0.284  0.77927
carb         -30.107      7.551  -3.987  0.00067 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.96 on 21 degrees of freedom
Multiple R-squared:  0.9549,    Adjusted R-squared:  0.9335
F-statistic: 44.51 on 10 and 21 DF,  p-value: 7.255e-12
```

From the result, the equation obtained will be:

$Y = -5.812 + 1.940X_1 + 15.389X_2 + 0.665X_3 + 8.812X_4 + 86.711X_5 – 12.974X_6 – 12.115X_7 – 7.914X_8 + 5.127X_9 – 30.107X_{10}$

As for part (b), the R solution has shown the significant variables, namely hp, wt and carb. Hence, the new equation will be formed only with the significant variables.

$Y = -5.812 + 0.665X_3 + 86.711X_5 – 30.107X_{10}$

The code to obtain the reduced regression model is as follows:

```r
model.red = lm(disp~hp+wt+carb,data=mtcars)
(summary.red=summary(model.red))
```

```
Call:
lm(formula = disp ~ hp + wt + carb, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-55.064 -27.878  -4.473  23.433  64.717

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -113.9825    23.1090  -4.932 3.34e-05 ***
hp             1.1970     0.1777   6.736 2.59e-07 ***
wt            76.7591     9.1138   8.422 3.69e-09 ***
carb         -27.6744     6.2779  -4.408 0.000139 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.02 on 28 degrees of freedom
Multiple R-squared:  0.9194,    Adjusted R-squared:  0.9108
F-statistic: 106.5 on 3 and 28 DF,  p-value: 2.033e-15
```

Comparison to the conventional approach

The approach that I am personally used to doing is the manual way. Compared to R, this is much more inefficient, and the model cannot be immediately recalculated based on the hypothesis test to see which variable is significant. With R, regression analysis becomes much more efficient.