**ISL Assignment 1-Q3**

**Consider again the auto.csv dataset from Q-1.**

i. Perform linear regression on **mpg** as the response with the following predictors: **cylinders, displacement, weight, acceleration, year, origin.**

**Answer:**

```
> Auto_rev.mod2=lm(mpg~cylinders+displacement+weight+acceleration+year
+origin,data=Auto_rev)
> Auto_rev.mod2
```

**Output:**

```
Call:
lm(formula = mpg ~ cylinders + displacement + weight + acceleration +
year + origin, data = Auto_rev)

Coefficients:
(Intercept)     cylinders  displacement        weight  acceleration
-19.743798     -0.444697      0.017186     -0.006838      0.155664
year           origin
0.764716       1.346033
```

ii. **Provide the summary report.**

**Answer:**

```
> summary(Auto_rev.mod2)

Call:
lm(formula = mpg ~ cylinders + displacement + weight + acceleration +
 year + origin, data = Auto_rev)

Residuals:
Min      1Q  Median      3Q     Max
-9.5640 -2.1692 -0.0382  1.8196 13.0720

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.974e+01  4.168e+00  -4.737 3.06e-06 ***
cylinders    -4.447e-01  3.211e-01  -1.385   0.1668
displacement  1.719e-02  7.189e-03   2.390   0.0173 *
weight       -6.838e-03  5.812e-04 -11.767  < 2e-16 ***
acceleration  1.557e-01  7.777e-02   2.002   0.0460 *
year          7.647e-01  4.973e-02  15.378  < 2e-16 ***
origin        1.346e+00  2.706e-01   4.975 9.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.33 on 385 degrees of freedom
Multiple R-squared:  0.8208,   Adjusted R-squared:  0.818
F-statistic: 293.9 on 6 and 385 DF,  p-value: < 2.2e-16
```

iii. **Which predictors do not have influence on** mpg **(in statistical sense) and why?**

**Answer: The influence of predictors on response(mpg) can be defined by using the p-values associated to each predictor**.

If p- value is very small, then we can infer that there comes alternative hypothesis i.e., there is a relationship between the predictor and response.

```
> Auto_rev.mod3=lm(mpg~cylinders+displacement+horsepower+weight+accele
ration+year+origin,data=Auto_rev)
> summary(Auto_rev.mod3)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin, data = Auto_rev)

Residuals:
Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,   Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16

In this scenario from the summary above, predictors “cylinders”, “hors
epower” and “acceleration” and having p-value >0.05 .Thus, these predi
ctors are not statistically significant with mpg.
```

iv. **Re-run the model with the remaining subset of predictors that have influence on** mpg**. Provide the summary report and comment on how this differs from part-iii in terms p-value, $R^2$ etc.**

**Answer:** Predictors displacement, weight, year and origin have been having statistically significant relationship with the response mpg. Below is the summary report for the predictors that has influence**.**

```
> Auto_rev.mod4=lm(mpg~displacement+weight+year+origin,data=Auto_rev)
> summary(Auto_rev.mod4)

Call:
lm(formula = mpg ~ displacement + weight + year + origin, data = Auto_
rev)
Residuals:
 Min      1Q  Median      3Q     Max
-9.8102 -2.1129 -0.0388  1.7725 13.2085

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.861e+01  4.028e+00  -4.620 5.25e-06 ***
displacement  5.588e-03  4.768e-03   1.172    0.242
weight       -6.575e-03  5.571e-04 -11.802  < 2e-16 ***
year          7.714e-01  4.981e-02  15.486  < 2e-16 ***
origin        1.226e+00  2.670e-01   4.593 5.92e-06 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.346 on 387 degrees of freedom
Multiple R-squared:  0.8181,   Adjusted R-squared:  0.8162
F-statistic: 435.1 on 4 and 387 DF,  p-value: < 2.2e-16
```

**Comment on the o/p:**

**Having the comparison between both the models have same p-value <**2.2e-16. The p value is very small for both the models. Thus, according to alternative hypothesis, we can infer that there is an association between the predictor and the response. Such that we can reject the null hypothesis for both the models.

$R^2$ is a statistical measure for to determine how close the data are to the fitted regression line. Generally, the higher the R squared the better the model fits your data.

In this scenario

Non-influence predictors model has R-squared 0.8215 (~82%)

Influence predictors model has R-squared 0.8181(~82%)

Hence, we can conclude that both the models are similar in terms of p value and $R^{2.}$

v. **(*must for Graduate students, optional for Undergraduate students for bonus points*)**
**Analyze by considering *all* the predictors in the dataset and how it influences** mpg **as response**.
**Answer:**

```
> Auto_rev.mod5=lm(mpg~cylinders+displacement+horsepower+weight+accele
ration+year+origin,data=Auto_rev)
> summary(Auto_rev.mod5)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
acceleration + year + origin, data = Auto_rev)

Residuals:
Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,   Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

**Analyzation:** Hence, we can analyze that the p-value is very small, so we can infer that the above data model has relationship among predictors and response by excluding null hypothesis. And by seeing the R-square 0.8215 value we can analyze that there is a strong relationship ($R^2$ >=0.5 is fairly a good cor relation).