NPTEL DATA SCIENCE FOR ENGINEERS

ASSIGNMENT 7

Based on the information given below answer the questions 1 to 5:

Description: Auto.csv dataset contains the details about the different parts of the cars. Objective of this problem is to predict mpg (mile per gallon) using the other predictors given in the dataset.

Variables	**Description**
mpg	miles per gallon
cylinders	Number of cylinders between 4 and 8
displacement	Engine displacement (cu. inches)
horsepower	Engine horsepower
weight	Vehicle weight (lbs.)
acceleration	Time to accelerate from 0 to 60 mph (sec)

1. The total number of missing values in the data frame is.

Solution: b

```
> #Read the dataset after setting the working directory
> data=read.csv("Auto.csv")
> sum(is.na(data))
[1] 0
```

2. The Pearson's correlation coefficient between **mpg** & **acceleration** is (rounded off to two decimal places): -

Solution: a

```
> round(cor(data$mpg,data$acceleration),2)
[1] 0.42
```

Build a linear regression model **"lr_model"** using all the variables in the data. Questions 3, 4 and 5 below are based on the **"lr_model".**

3. What is the value of adjusted R-Squared for **"lr_model"**?

Solution: c

```
> lr_model<- lm(mpg~., data = data)
> summary(lr_model)

Call:
lm(formula = mpg ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5816  -2.8618  -0.3404   2.2438  16.3416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower   -4.526e-02  1.666e-02  -2.716   0.0069 **
weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

4. The coefficient of the variable ‘displacement’ is:

Solution: c

```
> lr_model<- lm(mpg~., data = data)
> summary(lr_model)

Call:
lm(formula = mpg ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5816  -2.8618  -0.3404   2.2438  16.3416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders    -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower   -4.526e-02  1.666e-02  -2.716   0.0069 **
weight       -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

5. Which of the variables is not significant in “lr_model”?

Solution: a

```
> lr_model<- lm(mpg~., data = data)
> summary(lr_model)

Call:
lm(formula = mpg ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5816  -2.8618  -0.3404   2.2438  16.3416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.626e+01  2.669e+00  17.331   <2e-16 ***
cylinders   -3.979e-01  4.105e-01  -0.969   0.3330
displacement -8.313e-05  9.072e-03  -0.009   0.9927
horsepower  -4.526e-02  1.666e-02  -2.716   0.0069 **
weight      -5.187e-03  8.167e-04  -6.351    6e-10 ***
acceleration -2.910e-02  1.258e-01  -0.231   0.8171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 386 degrees of freedom
Multiple R-squared:  0.7077,    Adjusted R-squared:  0.7039
F-statistic: 186.9 on 5 and 386 DF,  p-value: < 2.2e-16
```

Consider the following confusion matrix to answer Q6 and Q7.

		Actual	
		Accept	Reject
Predicted	Accept	15	5
	Reject	1	5

6. The accuracy of the model is (rounded off to two decimal places): -

Solution: c

$N = TP + FP + TN + FN = 15 + 5 + 1 + 5 = 26$

Accuracy=$\frac{TP+TN}{N}$ = $\frac{15+5}{26}$ = 20/26 = 0.769 ~0.77

7. The sensitivity pertaining to the given confusion matrix is (rounded off to two decimal places)

Solution: a

Sensitivity = $\frac{TP}{TP+FN} = \frac{15}{15+1} = \frac{15}{16} = 0.9375 \sim 0.94$

8. Which command is used to build a logistic regression model in R?

Solution: a

glm () is used to build a logistic regression model.

9. The Logistic regression tends to overfit when we have large number of independent variables present.

Solution: a

Yes, the logistic regression tends to overfit when we have large number of independent variables present.

10. An ROC curve is plotted between.

Solution: b

An ROC curve can be plotted between Sensitivity and (1 – Specificity) or True Positive Rate and False Positive Rate, to determine the better classifier that accurately predicts the classes.