Assignmetn3

Name: Ziling Huang

Username: zhu51

# question1

a)

training <- read.csv("training.csv")

head(training)

tail(training)

Table

Description automatically generated with medium confidence

b)

Chart, scatter chart

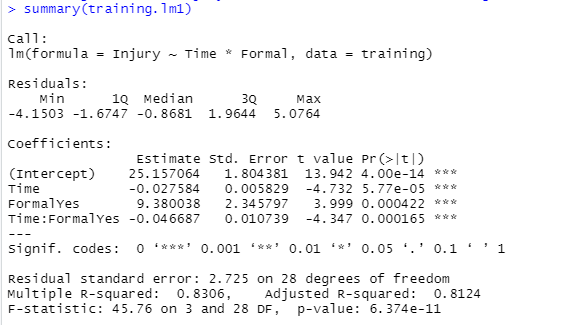
Description automatically generated

From the plot graph, we can see the formal is a factor with two levels, yes or no, which are red and black respectively. Overall, there is a decreasing and linear trend between injury and time. As time increases, the injury decreases.

Chart, box and whisker chart

Description automatically generated

From the box plot above, we can see the median of formal training is higher than the median of informal training. The lower quartile (25%) of formal training is higher than the upper quartile (75%) of informal training. Also, the data in formal training is more spreading out than the informal training, which means observations vary more. Overall, the box plot is suggesting that formal training is not as good as informal training for reducing injuries.

c)  


Text

Description automatically generated

When Formal = “No”, injury = 25.16 – 0.028 \* Time

When Formal = “Yes”, injury = 25.16 + 9.38 + (-0.028 – 0.047) \* Time

From our summary to check are different slopes needed, we can see the p-value of Time and Time:FormalYes are 5.77 \* 10^(-5) and 0.000422 respectively. They are below 0.05, we can say they are significant. There is a relationship between them and Injury.

From Anova, we can see the interaction between Time and Formal is significant as P-value is less than 0.05. But we notice that the p-value of Formal is greater than 0.05, which is not significant. But as Formal itself would affect the slope (the relationship between Injury and Time), therefore we need to keep Formal as well.

Overall, we can’t reduce the model anymore.

d)  
Chart, scatter chart

Description automatically generated

From Residuals vs Fitted graph we can see the residuals are nearly evenly spread around 0. However, the red line is not that fitted the dash line, there is a deviation. But as there are not too many observations, we can still say the model is good enough to fit the linearity.

Chart, scatter chart

Description automatically generated

From Normal Q-Q graph, we can see the observations are not fitted the dash line. There are some deviations between observations from Normal distribution. But the residuals are not too much, those observations are still nearly around and fitted the dash line, which we can say the model is good enough. The model meets the normality.

Chart, histogram

Description automatically generated

From Cook’s distance, we can see there are three observations which are 20, 25 and 31 away from our dataset, but as their Cook’s distance is not greater than 0.4, so we say they are not significantly influence our model.

e)

From the summary of our model, there are two equations for formal training and informal training.   
When Formal = “No”, injury = 25.16 – 0.028 \* Time

When Formal = “Yes”, injury = 25.16 + 9.38 + (-0.028 – 0.047) \* Time

We can say when the time increases by 1 unit, injury in formal training will decrease 0.075, while the injury in informal training will decrease 0.028. In 1 unit increase of time, the injury will decrease faster by 0.047 than the injury in informal training. We can also tell from the dot-plot graph, as the black dots(informal training) are flatter than the red dots(formal training).

f)  
When Formal = “No”, injury = 25.16 – 0.028 \* Time

When Formal = “Yes”, injury = 25.16 + 9.38 + (-0.028 – 0.047) \* Time

In the question, when the Formal = “No”, time = 300, predict injury score

Injury = 25.16 – 0.028 \* 300 = 16.88

So, the predicted injury score is 16.88

# Question2

a)

autompg <- read.csv("autompg.csv")

autompg$Cyl <- as.factor(autompg$Cyl)

b)

plot(MPG~Disp,

data=autompg, pch = 19, col = Cyl)

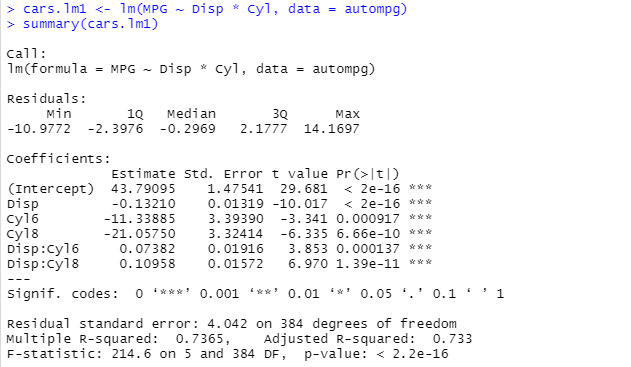
# black for the 4 cylinder cars, red for the 6 cylinder cars and green for the 8 cylinder cars

Chart, scatter chart

Description automatically generated

We can observe that there are three levels for cylinder are 4 cylinder, 6 cylinder and 8 cylinder, the colors are black, red and green respectively. If we ignore about the factor, we can observe that the relationship between displacement and MPG looks like a decreasing quadratic relationship. And if we include the factor, we found the lower the cylinder, the higher the MPG, but less the displacement. The higher the cylinder, the lower the MPG, but higher the displacement.

c)



Text

Description automatically generated

d)  
Chart, scatter chart

Description automatically generated

From the Residuals vs Fitted graph, we can see the residuals are evenly spreading around the zero, and the red line is fitted the zero dash line. We could say the model meets the linearity.

Chart, line chart

Description automatically generated

From the Normal Q-Q graph, we can see there are some deviations at the tails. But overall the residuals fit the dash line, which means there is not too much deviations from a Normal distribution. So we can say the model meets the normality

Chart, histogram

Description automatically generated

From Cook’s distance graph, we can see there are three significant points are away from our dataset which are 274, 328 and 354. As their cook’s distance is less than 0.4, we could say it doesn’t highly influence our model.

e)  
we can’t simplify our model further. As from our summary, we could notice that the p-value for the slope of the 4 cylinder, 6 cylinder and 8 cylinder are 2e-16, 0.000137 and 1.39e-11 respectively, which are all less than 0.05. That shows they are all significant, there is a relationship between the different cylinder and the displacement.

From the Anova, we can observe that the p-value of cylinder and the interaction between displacement and cylinder are all below 0.05 which tells they are significant to our model.

From the information above, we can say the cylinder is in a relationship between our model whether in each factor level or as a cylinder overall. We can’t simplify the model anymore.

f)  
From the dot-plot graph, we can see the higher the cylinder, the higher the displacement but lower the MPG. The lower the cylinder, the lower the displacement but the higher the MPG.

From the summary, we have the equations:

When 4 cylinder: MPG = 43.79 – 0.13 \* Displacement

When 6 cylinder: MPG = (43.79 – 11.33885) + (0.074 – 0.13) \* Displacement

When 8 cylinder: MPG = (43.79 – 21.05750) + (0.10958 – 0.13) \* Displacement

Code:

#question1

#a)

training <- read.csv("training.csv")

head(training)

tail(training)

#b)

training$Formal <- as.factor(training$Formal)

plot(Injury~Time,

data=training, pch = 19, col = Formal)

boxplot(Injury ~ Formal, data = training)

#c)

training.lm1 <- lm(Injury ~ Time \* Formal, data = training)

summary(training.lm1)

anova(training.lm1)

#d)

plot(training.lm1, which = 1)

plot(training.lm1, which = 2)

plot(training.lm1, which = 4)

#e)

#f)

#question2

#a)

autompg <- read.csv("autompg.csv")

autompg$Cyl <- as.factor(autompg$Cyl)

#b)

plot(MPG~Disp,

data=autompg, pch = 19, col = Cyl)

# black for the 4 cylinder cars, red for the 6 cylinder cars and

# green for the 8 cylinder cars

#c)

cars.lm1 <- lm(MPG ~ Disp \* Cyl, data = autompg)

summary(cars.lm1)

anova(cars.lm1)

#d)

plot(cars.lm1, which = 1)

plot(cars.lm1, which = 2)

plot(cars.lm1, which = 4)