Joseph Flynn

STA 4155

1. Let’s take MSRP($) as response variable and consider Wheelbase(in), Displacement(cu in), Bore(in)and Clearance(in) as potential predictors. Use scatterplots to see which variables can be appropriately used as predictors in simple linear regression.

The scatterplots for both Displacement vs. MSRP and Bore vs. MSRP show a moderate to strong positive correlation and can be used as predictors without any change. Unlike wheelbase and clearance.

plot(motor$Displacement,motor$MSRP,xlab="Displacement",ylab="MSRP",main =

"Displacement vs. MSRP")

Chart, scatter chart

Description automatically generated

plot(motor$Bore,motor$MSRP,xlab="Bore",ylab="MSRP",main = "Bore vs. MSRP")

Chart, scatter chart

Description automatically generated

Wheelbase and clearance each have a curve that is similar in shape to y = 2x. The way to linearize this would normally be to take the log value of y. However, this can’t currently be done because we want the same y for all predictors. So, we instead put the predictor into the exponent position with of a base. After doing, we find the predictors 1.2wheelbase and 1.15Clearance create a moderate linear pattern with MSRP.

A picture containing object, clock

Description automatically generated

par(mfrow = c(1,2))

plot(motor$Wheelbase,motor$MSRP,xlab= "Wheel Base", ylab = "MSRP", main =

"Wheel Base vs. MSRP") #linerize

plot(1.20^motor$Wheelbase,motor$MSRP, xlab= "1.2^Wheel Base", ylab = "MSRP",

main = "1.2^Wheel Base vs. MSRP")

Chart, scatter chart

Description automatically generated

par(mfrow = c(1,2))

plot(motor$Clearance,motor$MSRP,xlab="Clearance",ylab = "MSRP",main =

"Clearance vs. MSRP") #linerize

plot(1.15^motor$Clearance,motor$MSRP,xlab="1.15^Clearance",ylab = "MSRP",main

= "1.15^Clearance vs. MSRP")

Chart, scatter chart

Description automatically generated

From here, the possible predictors for MSRP are Displacement, Bore, 1.2wheelbase, and 1.15Clearance.

2. Build a multiple regression model for MSRP using Displacement and Boreas predictors. Write down the fitted model. Report R2 and adjusted R2. Interpret the coefficients for Displacement and Bore.

After setting up a regression model, the fitted model Predicted MSRP = 423.025 + 6.722(Displacement) + 38.915(Bore) is produced from the prediction values. From the summary of the model an R2 = 0.7566 and Adjusted R2 = 0.7512 is found that .7512 of variation is attributed to the model after taking account for the number of predictors. Additionally, the displacement coefficent can be interpretted as when displace increases by 1, predicted MSRP increases by 6.722. For the Bore coefficent, and increase of 1 for Bore causes MSRP to increase by 38.915.

model.1 = lm(MSRP ~ Displacement+Bore, data = motor)

summary(model.1)

Text, table

Description automatically generated

3. Check the model you fitted in the previous question to see if it satisfies the assumptions as required inmultiple regression.

To check the model requirements, a residual plot and qq-plot are produced.

plot(model.1$residuals)

abline(0,0)

A picture containing scatter chart

Description automatically generated

From the residual plot, there is no pattern; meaning the Linearity Assumption is met. There aren’t any outliers either in the plot. The Equal Variance Assumption is also satisfied because the residuals have equal spread. Next would be to check normality with a histogram and Q-Q plot.

par(mfrow = c(1,2))

hist(model.1$residuals, main ='', xlab ='Residuals')

qqnorm(model.1$residuals)

qqline(model.1$residuals)

Chart, histogram

Description automatically generated

The Q-Q plot and histogram show that the Normality Assumption is failed because the Q-Q plots stray from the line, showing weight in the tails. While the histogram shows a distribution uncentered from zero with a noticable right skew.

4. Conduct a test to see if the fitted multiple regression model is statistically useful. If useful, find the predictors that make significant contributions to the MSRPi n the model. Explain.

From the summary, shown below, the model is shown to be statistically useful because the p-value for the F-statistic is < 2.2e-16, which means that the model has atleast one predictor that is not equal to zero and is useful to predict MSRP. For the F-test null hypothesis is rejected at a .05 alpha level. However, Displacement seems to be the only signifcant predictor because it is the predictor with a probability less than the alpha level of .05. Meaning that Bore doesn’t contribute much to the model and maybe equivilant to zero.

summary(model.1)

#F-statistic: 139.9 on 2 and 90 DF, p-value: < 2.2e-16Text, table

Description automatically generated

5. Suppose we are not satisfied with theR2 given by the current model. Please propose a new multiple regression model in order to improve R2. Compare the new model to the current one with respect to their R2, coefficient estimates and hypothesis tests. Don’t forget to check assumptions of the new model for its validity. (Hint: We have two potential predictorsWheelbase andClearancein the pool. Think about howto use them to improve the model.)

The model we created before was still missing the linearized wheelbase and clearance. So, lets create a model with all the prediction variables. The summary of which is shown below. From this, we can see that Displacement, linearized wheelbase, and linearized clearance have a p-value less than the alpha of .05 and even less than an alpha of .01. So, we can eliminate Bore from the model.

wheel\_lin = 1.25^motor$Wheelbase

clear\_lin = 1.15^motor$Clearance

model.2 = lm(MSRP ~ Displacement+Bore+ wheel\_lin+ clear\_lin, data = motor)

summary(model.2)

Table

Description automatically generated

The next model eliminates Bore. It also eliminates wheelbase linearized because the residual plot eliminates the equal variation assumption. So, the final model involves Displacement and the linearized clearance. This model is shown below and has an adjusted R2 of 0.8938 and has all predictors below an alpha of 0.001.

model.3 = lm(MSRP ~ Displacement+clear\_lin, data = motor)

summary(model.3)

Text

Description automatically generated

From this the Equal variance assumption is shown by the equal spread of residuals. The Linearity assumption is demonstrated by no pattern being in the residuals. In addition, there doesn’t seem to be any outliers. Then the Q-Q plot hugs fairly close to the line and the histogram is fairly normal. Thus passing the normality assumption.

plot(model.3$residuals)

abline(0,0)

Chart, scatter chart

Description automatically generated

par(mfrow =c(1,2))

hist(model.3$residuals, main ='', xlab ='Residuals')

qqnorm(model.3$residuals)

qqline(model.3$residuals)

Chart, histogram

Description automatically generated

The final model passed the assumptions and is Predicted MSRP = 8.6614(Displacement) + 458.1690(1.15Clearance).