Stat 415 Regression: Homework 2 (Multiple Regression)

Frankie Tyndall

2022-07-24

## Gas Mileage Study

A researcher is interested in developing a model that describes the gas mileage, measured in miles per gallon (mpg), of automobiles. Based on input from an engineer, she decides that the explanatory variables might be engine size (liters), curb weight (pounds) and horsepower.From a random sample of 13 automobiles, she obtains the following data.

explanatory variables: engine size, curb weight, horsepower response variable: gas mileage (MilesperGallon)

## Gas Mileage Dataset

tribble(~EngineSize, ~CurbWeight, ~Horsepower, ~MilesperGallon,  
 2.4, 3289, 177, 24,  
 2.4, 3263, 158, 25,  
 2.5, 3230, 170, 24,  
 3.5, 3580, 272, 22,  
 2.8, 3175, 255, 18,  
 3.5, 3643, 263, 22,  
 3.5, 3497, 306, 20,  
 3.0, 3340, 230, 21,  
 3.6, 3861, 263, 19,  
 2.4, 3287, 173, 24,  
 3.3, 3629, 234, 21,  
 2.5, 3270, 170, 22,  
 3.5, 3292, 270, 22  
) -> V  
 V

## # A tibble: 13 × 4  
## EngineSize CurbWeight Horsepower MilesperGallon  
## <dbl> <dbl> <dbl> <dbl>  
## 1 2.4 3289 177 24  
## 2 2.4 3263 158 25  
## 3 2.5 3230 170 24  
## 4 3.5 3580 272 22  
## 5 2.8 3175 255 18  
## 6 3.5 3643 263 22  
## 7 3.5 3497 306 20  
## 8 3 3340 230 21  
## 9 3.6 3861 263 19  
## 10 2.4 3287 173 24  
## 11 3.3 3629 234 21  
## 12 2.5 3270 170 22  
## 13 3.5 3292 270 22

## Use and show R code that will produce a correlation matrix.

cor(V)

## EngineSize CurbWeight Horsepower MilesperGallon  
## EngineSize 1.0000000 0.7635126 0.9204991 -0.5966553  
## CurbWeight 0.7635126 1.0000000 0.5579994 -0.3899979  
## Horsepower 0.9204991 0.5579994 1.0000000 -0.7349600  
## MilesperGallon -0.5966553 -0.3899979 -0.7349600 1.0000000

## Indicate if the correlation matrix shows multicollinearity problems with pairs of explanatory variables. The variable Miles per Gallon is the response variable

Using the correlation matrix multicollinearity is definitely found between pairs of the explanatory variables. The pair horsepower and engine size has a high correlation of 0.9204991. Engine size and Curb weight also have a relatively strong correlation. These may indicate that some of the predictor variables are not needed in the model.

## Indicate the explanatory variable(s) that the response variable is strongly correlated with.

All of the explanatory variables have a negative correlation with the response variable. Horsepower has the strongest correlation with the reponse variable with a correlation coefficient of -0.7349600.

## Use and show R code that will produce a full multiple regression model, that uses all explanatory variables. Write the full model that shows the coefficients for all explanatory variables and the intercept.

V\_model <- lm(MilesperGallon~EngineSize + CurbWeight + Horsepower, V)  
V\_model

##   
## Call:  
## lm(formula = MilesperGallon ~ EngineSize + CurbWeight + Horsepower,   
## data = V)  
##   
## Coefficients:  
## (Intercept) EngineSize CurbWeight Horsepower   
## 35.826758 4.803063 -0.003945 -0.065831

The Full Model:

MilesPerGallon = 35.826758 + 4.803063*EngineSize - 0.003945*CurbWeight - 0.065831\*Horsepower

## Use and show R code that will produce all important summary statistics for your model.

V\_model <- lm(MilesperGallon~EngineSize + CurbWeight + Horsepower, V)  
summary(V\_model)

##   
## Call:  
## lm(formula = MilesperGallon ~ EngineSize + CurbWeight + Horsepower,   
## data = V)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9635 -0.9565 0.1233 1.0472 1.3911   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.826758 9.211631 3.889 0.00368 \*\*  
## EngineSize 4.803063 3.360832 1.429 0.18674   
## CurbWeight -0.003945 0.003847 -1.026 0.33187   
## Horsepower -0.065831 0.026360 -2.497 0.03401 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.466 on 9 degrees of freedom  
## Multiple R-squared: 0.6257, Adjusted R-squared: 0.5009   
## F-statistic: 5.015 on 3 and 9 DF, p-value: 0.02585

## Which explanatory variable(s) is/are significant at the level of .05?

Horsepower is the only explanatory variable with a p value less than 0.05 with a p value of 0.03401. Horsepower is the only explanatory variable that is significant in the model.

## Use your book or an internet source and describe the difference between R2 and adjusted R2. (Three or four sentence)

The difference between R^2 and adjusted R^2 is that takes into account all explanatory variables, whereas the adjusted R squared only factors in explanatory variables that truly have an impact on the model’s performance. Both are ways to study the proportion of variability in the response variable explained by the explanatory variable/s but adjust R squared only takes into account the variables that make an impact on the model to best explain the variation in the response variable. By taking into account the impact of additional independent factors that have a tendency to skew the outcomes of R-squared measurements, adjusted R-squared increases precision and dependability.

## What is the proportion of the variability in Milespergallon that is explained by the full model regression equation?

The multiple R-squared: 0.6257

62.57% of the variation in Milespergallon is explained by the full model regression equation.

## What is the proportion of the variability in Milespergallon that is explained by the full model regression equation, only using explanatory variables that have significant impact on the dependent variable.

The Adjusted R-squared: 0.5009

50.09% of the variablity in Milespergallon is explained by the full model regression equation, only using predictor variables that have significant impact on the dependent variable.

## Use and show R code that produces a model without the explanatory variable with the highest p value. Write the new model that indicates the response variable, the explanatory coefficients and the intercept.

removed\_V\_model <- lm(MilesperGallon~EngineSize + Horsepower, V)  
summary(removed\_V\_model)

##   
## Call:  
## lm(formula = MilesperGallon ~ EngineSize + Horsepower, data = V)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3007 -1.1158 0.7471 0.9924 1.3668   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 26.80495 2.74010 9.782 1.94e-06 \*\*\*  
## EngineSize 2.14695 2.14743 1.000 0.3410   
## Horsepower -0.05032 0.02164 -2.325 0.0424 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.47 on 10 degrees of freedom  
## Multiple R-squared: 0.582, Adjusted R-squared: 0.4983   
## F-statistic: 6.96 on 2 and 10 DF, p-value: 0.01277

## Explain why the null hypothesis that the population coefficient for Engine Size is equal to zero is not to be rejected.

The response variable, Engine Size, has a p value of 0.3410 in the model with the highest p value response variable removed. This p value is larger than the threshold value of 0.05 which means that it is not significant so you would not reject that the population coefficient is equal to zero.

## Which model, the one with three explanatory variables or the one with two explanatory variables is the better model for predicting Miles per gallon? Justify your answer in three of four sentences.

The model with three explanatory variables seem to be the better fit than the model with just two variables. Firstly, the model with three explanatory variables have a multiple R-squared of 0.6257 and an adjusted R-squared of 0.5009, the model with only two have smaller multiple R squared and adjusted R-squared. The model with three have higher R-squared values which is an indicator a better model. In the model with three, the p values of the predictor variable were smaller, this also indicates that it was the better model.