EUROPA-
UNIVERSITÄT
VIADRINA
FRANKFURT
(ODER)

The Factors That Impact the Price of Skoda Fabia Cars

A Seminar Paper submitted for the course
“Introduction the Statistics and Data Science”
Professor: Dr. Dmytro Ivasiuk

Recep Göktuğ Şengün
Matriculation Number: 128272

TABLE OF CONTENTS

1. INTRODUCTION 3
2. THEORY 3
3. DATASET 3
    3.1. Variables 4
4. MEAN AND VARIANCE 5
    4.1. Analysis of Age and Price 5
    4.2. Analysis of Mileage and Price 6
5. MULTIPLE REGRESSION MODEL 7
    5.1. Estimation of the Model 7
    5.2. Parameter Significance and the Goodness-of-Fit 7
    5.3. Stepwise Selection 8
    5.4. Model Visualization 9
    5.5. How does the model work? 9
    5.6. Confidence Intervals of Residuals 10
    5.7. Predicted Price and the Actual Price 10
6. CONCLUSION 11
7. REFERENCES 12

1. INTRODUCTION

The Skoda - Fabia is a popular compact car that has been widely praised for its practicality, reliability, and affordability. Over the years, it has become a sought-after option for those in the market for a used car, as it offers a good balance of comfort, performance, and affordability. When it comes to the used car market, there are several parameters that can impact the price of a Skoda Fabia. These factors include the age of the car, the number of miles it has been driven, its condition, and the features it comes equipped with.
In this analysis, I will take a closer look at the different parameters that can affect the price of a used Skoda Fabia and explore how they can impact the overall value of the car. Understanding these factors can be useful for anyone who is looking to purchase a used Skoda Fabia and wants to get the best value for their money.
The data which is used for this analysis is gathered from "www.mobile.de" , It consists of observations of cars whose first registration date is between 2017 and 2022, type of fuel is petrol and type of transmission is automatic.

2. THEORY

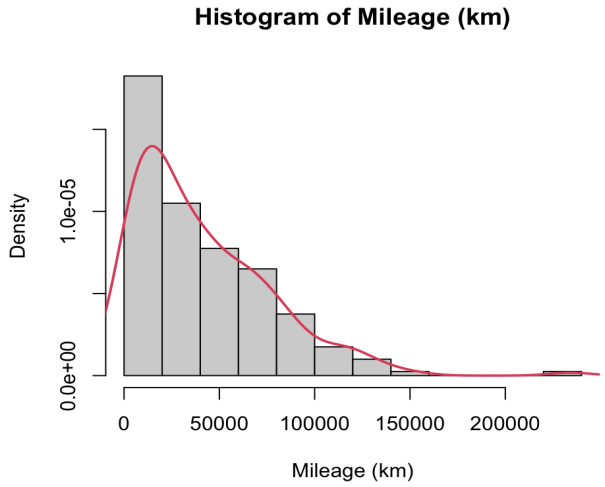
According to mainstream economic theory the prices of used cars have a close relation to the expected value a car can provide its owner during the remainder of the car's lifetime. As a car grows older, it is expected that the remaining lifetime decreases. The standard model thus predicts that, when all other things are equal, older cars will have lower prices on the used car market. Also, a car of a given age with high mileage is likely to break down earlier than a car of the same age. Hence, mainstream economics predicts that cars with higher mileage have lower prices considering all other things are equal. Additionally, an engine's power and cubic capacity is believed to affect the price of a car.
In this study I aim at finding out which of the factors have the highest influence on the car's price. This will be done analyzing the data acquired from mobile.de, one of the largest automotive vehicles online markets in Europe. Not only is this important to know for people planning on buying a car, but also to understand the real value of a car.

3. DATASET

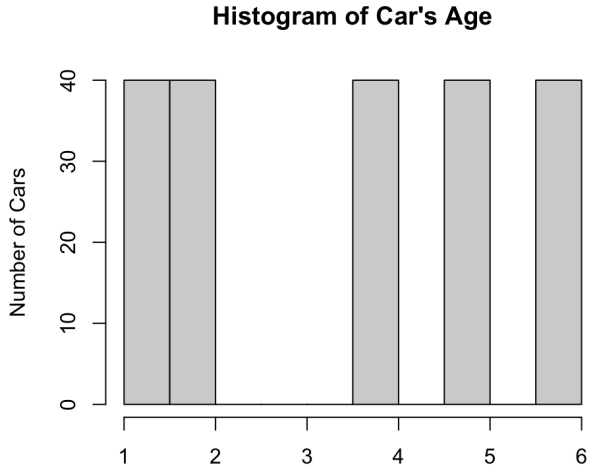
I collected the information of about 200 Skoda - Fabia cars which are listed in January 2023 on "www.mobile.de", which is one of Europe's largest online vehicle markets. On the website, people can list and sell their cars and buy new or used ones. Since the company is not involved in the transactions, it can be assumed that prices are determined according to supply and demand, therefore, it is a free marketplace. In order to search for a certain car or a model, prospective buyers can use the search tab of the website. In order to reach desired car type I filled the tabs as follows:
Make: "Skoda",
Model: "Fabia",
Type and condition: "Used",
Country: "Germany",
Fuel Type: "Petrol",
Transmission: "Automatic transmission"

First registration: “2017- Current”.
I extracted “date”, “mileage(km)”, “engine’s power(kW)” and “engine’s cubic capacity” to use as my independent variables, and to analyze their effects on it, I took the dependent variable “price(€)”. Then, I converted the registration date (date) variable into the “age” variable by subtracting the registration date from the current year.

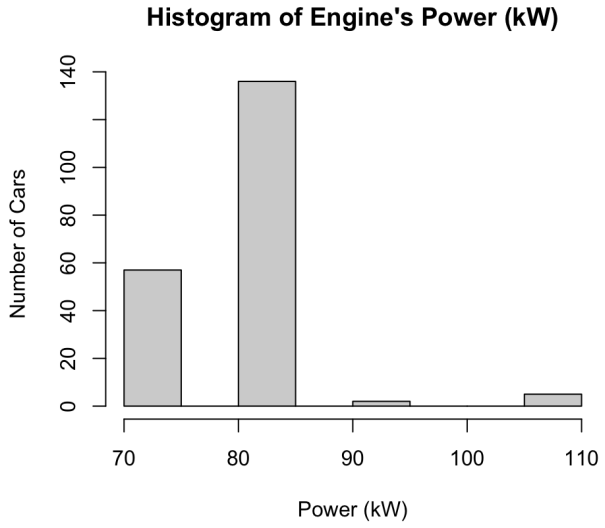
3.1. Variables



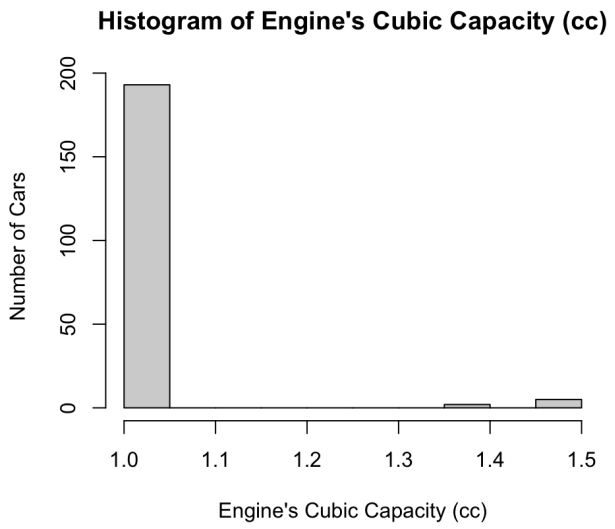
The histogram shows the number of cars within different mileage intervals. The number of cars within the range of 0 - 20.000 km is the highest in the dataset. And the number of cars decreases with each subsequent interval. So there are less cars listed with higher mileage.



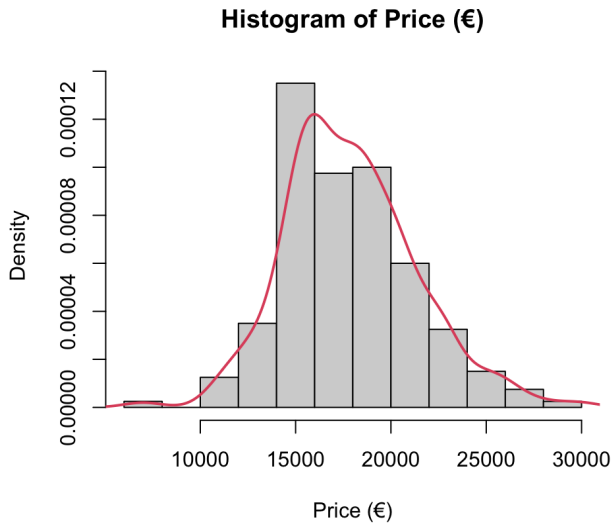
The Age variable consists of 40 observations of each age group from 1-year-old up to 6-year-old. Since there was not enough data for 3-year-old cars, they are not included.



The data consists of 57 - 70 kW;
136 - 81 kW;
2 - 92 kW;
5 - 110 kW cars.



The histogram shows:
193 - 1.0 cc;
2 - 1.4 cc;
5 - 1.5 cc engines.

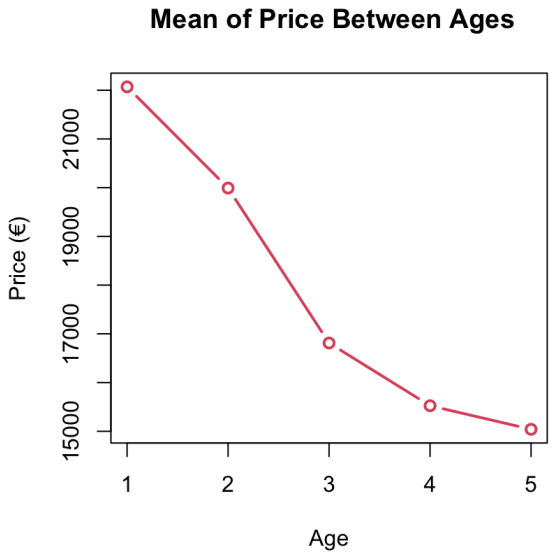


The dependent variable “Price”. Since it is more or less bell shaped, it shows an approximate normal distribution.

4. MEAN AND VARIANCE

4.1. Analysis of Age and Price

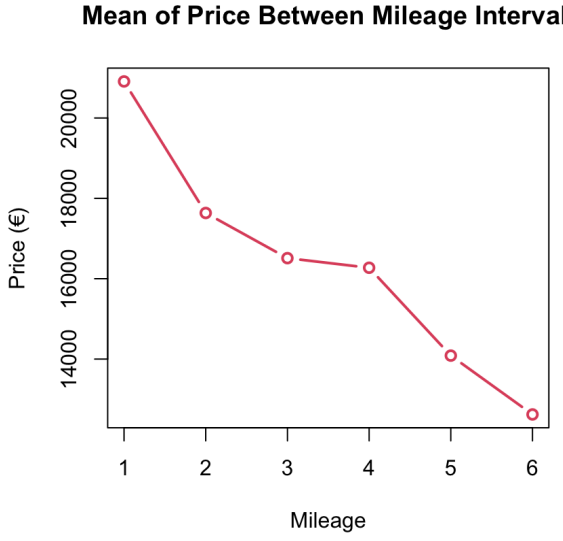
In order to analyze and understand the effects of age, I calculated the mean and variance of car prices for different ages. In the graph below, it can easily be seen that as cars grow older, the mean price of the cars decrease. This information gives us the direct impact of age on price. Thus it can be said that the younger cars have higher prices.

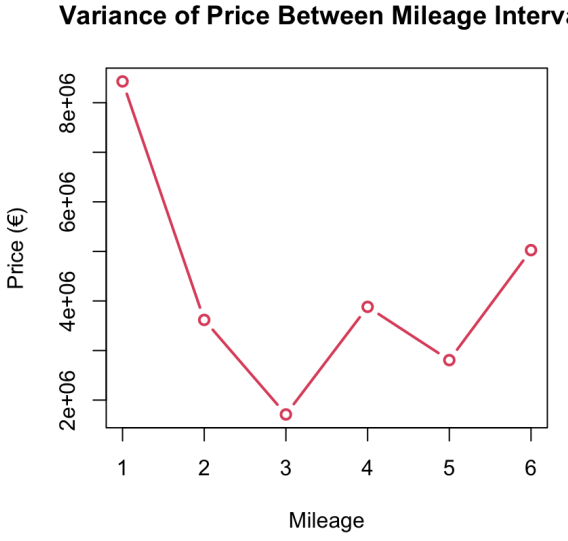


When looking at the variance, it is seen that it gets less for the higher ages except for age 4. This might be because of other measures such as mileage, since the age variable is not the only one to affect price. However, this could be simply because of outliers.

4.2. Analysis of Mileage and Price

After looking at the effects of age on price, I analyzed another important factor “Mileage”. In order to get a more understanding view, I divided the continuous mileage variable into categories. The first category consists of cars with 0 to 20.000 km; second 20.000 to 40.000 km; third 40.000 to 60.000 km; fourth 60.000 to 80.000 km; fifth 80.000 to 100.000 and last category, cars with 100.000 and higher mileage.





On the graphs above, it is seen the mean and variance in different mileage categories. When looking at the mean graph it is seen that cars with higher mileage have lower prices.

5. MULTIPLE REGRESSION MODEL

5.1. Estimation of the Model

For estimating a multiple regression model, I looked at the correlations of independent variables with the dependent variable using cor() function. The Age variable shows the strongest correlation with Price with -0.78 and mileage shows the second strongest correlation with Price with -0.75. Although other variables such as engine_cc(0.33 correlation) and power-kW(-0.15 correlation) do not show significant correlation, I included all variables to the model.
In order to perform a multiple linear regression analysis and check the results, I needed to run two lines of code. The first line of code makes the linear model, and the second line prints out the summary of the model:

model <- lm(Price ~ km + kW + age + engine_cc , data = fabia)
summary(model)

```
Call:
lm(formula = Price ~ km + kW + age + engine_cc, data = fabia)

Residuals:
    Min      1Q  Median      3Q     Max
-3820.9  -813.5  -124.1   860.5  4257.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.107e+04  1.237e+03   8.951 2.77e-16 ***
km          -4.362e-02  3.524e-03 -12.380  < 2e-16 ***
kW           2.139e+02  2.712e+01   7.888 2.14e-13 ***
age         -1.269e+03  8.289e+01 -15.310  < 2e-16 ***
engine_cc   -3.593e+03  2.059e+03  -1.744   0.0827 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1417 on 195 degrees of freedom
Multiple R-squared:  0.8353,    Adjusted R-squared:  0.8319
F-statistic: 247.2 on 4 and 195 DF,  p-value: < 2.2e-16
```

5.2. Parameter Significance and the Goodness-of-Fit

In order to understand the quality of a regression model, an R-squared measure is used. R-squared is a value between 0 and 1. If the value is close to "1", it shows a strong relationship and a better fit. The value of R-squared can be reached simply by using summary statistics. The summary statistics shows the R-squared along with other statistics. From the Adjusted R-squared section, the R-squared value is reached. As it is seen, the R-squared for the model is "0.8319" when all of the variables in the data are included. By looking at the summary statistics of the first model it can be said that the model is 83.19% good, in other words, there is a strong relationship between the dependent variable and the independent variables.

5.3 Stepwise Selection

I have included all the variables of the data in the first model. However, I can use other models and even get a better model by excluding some of them. One of the ways to determine which variables to include to the model is Adjusted R-squared. Although it is useful, it has a disadvantage. The disadvantage is that the R-squared score increases as the number of variables increases, even though there is not much relationship between the dependent variable and the independent variable.
A better approach for deciding on a better model is "Akaike's Information Criterion (AIC)". It is a better way of measuring because it takes the number of variables into consideration and avoids the effects of unrelated variables. In order to find the better variable to include in the model, the "AIC" score can be checked.

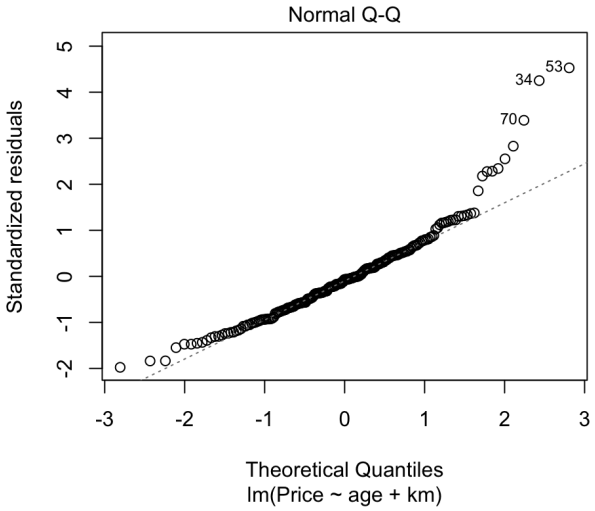
```
> stepAIC(model)
Start:  AIC=2907.42
Price ~ km + kW + age + engine_cc

            Df Sum of Sq       RSS    AIC
<none>                    391466001 2907.4
- engine_cc  1   6109171 397575172 2908.5
- kW         1 124912795 516378796 2960.8
- km         1 307659514 699125516 3021.4
- age        1 470524084 861990085 3063.3

Call:
lm(formula = Price ~ km + kW + age + engine_cc, data = fabia)

Coefficients:
(Intercept)           km           kW          age    engine_cc
  1.107e+04   -4.362e-02    2.139e+02   -1.269e+03   -3.593e+03
```

When I look at the outcome, I see our AIC score for the first model is "2907.42" when all variables are included. Then I look at each variable and notice that the engine's cubic capacity "engine_cc" has the lowest AIC score. Since I know that the variable with lower AIC score should be excluded in order to get a better model, I created a new model without "engine_cc".
Then, with the new model, the kW variable has the lowest AIC score, thus I remove it from the model. Lastly, the "age" is the most important variable according to AIC score.

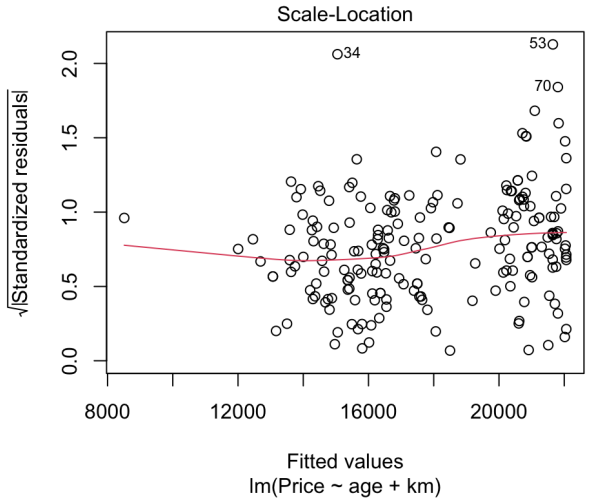
So the final model is:

model <- lm(Price ~ age + km) with "3008.37" AIC score.

5.4 Model Visualization



From the Q-Q plot, it can be seen that the residuals of the model are approximately linear, thus it supports that the errors are normally distributed.



The Scale-Location plot should look random and show no patterns. From the plot it is seen that it is random.

5.5 How does the model work?

The estimated regression equation is an equation constructed to model the relationship between dependent and independent variables.
The multiple linear regression equation that I created consists of 1 dependent variable (Price) and 2 independent variables (Age, Mileage).
In order to understand how the model works, the coef() function can be used. The function gives the intercept and the slopes of the model as a result.

```
> round(coef(model3), 2)
(Intercept)          km         age
   23007.68       -0.04     -944.73
```

As it is seen from the image, to predict new observations we multiply km value with “-0.04”, age value with “-944.73” and get the total of these values along with the intercept (23007.68). This gives us the estimated result.
It can be done easily by only using the predict() function in R. The predict() function makes these calculations automatically.

5.6 Confidence Intervals of Residuals

The residual for each observation is the distance between predicted values of y (dependent variable) and observed(real) values of y. So:
Residual = actual y value − predicted y value
In order to reach the residuals for the estimated model, I created a "predicted_price" column where I calculated the predicted price for the model and finally created a "residuals" field by subtracting fitted values from the actual "Price" field.

Then, the 95% confidence intervals for the residuals of the model are calculated with the given formula:

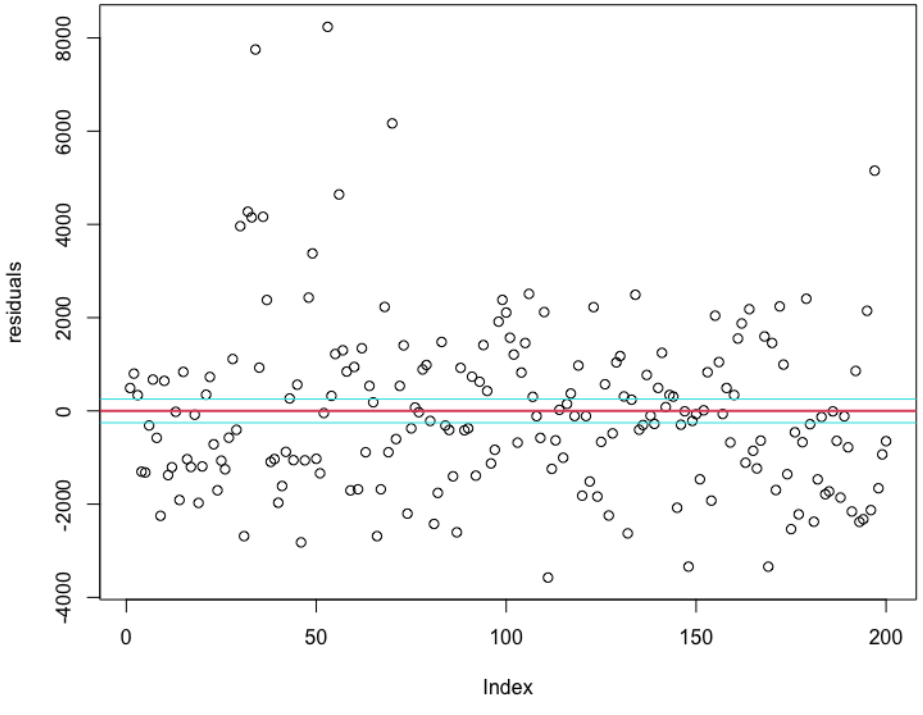
$$\left[\bar{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S_X}{\sqrt{n}}, \bar{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S_X}{\sqrt{n}}\right]$$

As a result of the calculations the residuals confidence intervals are:

- Lower Bound = -253.6118,
- Upper Bound = 253.6118.

5.7 Predicted Price and the Actual Price

The predicted car prices and the actual values can be compared by looking at the graph below. The graph shows the residuals and their 95% confidence intervals.
The red line in the middle where residual value is 0 shows the real price.
The blue lines show the 95% confidence intervals for residuals.



6. CONCLUSION

I studied the price-setting behavior in a competitive market for used cars and provided empirical evidence for information processing. I used the data from one of Germany's largest online marketplaces for vehicles. I analyzed the price differences and abnormalities of Skoda-Fabia cars with a focus on different registration dates, different mileages, different engine cubic capacities and power.

The results show that as in theory, in the used car market, the registration date and mileage of a vehicle play a significant role in determining its price. A newer car with lower mileage is typically more valuable and commands a higher price compared to an older car with higher mileage. This is due to the perception that a newer car has been less used and is therefore in better condition. Similarly, cars with lower mileage are seen as having been driven less and are therefore less likely to have incurred wear and tear.

On the other hand, cars that are older and have higher mileage are typically priced lower as they have a higher risk of incurring maintenance and repair costs. These factors are important to consider when determining the value of a used car and can greatly impact its price.

In conclusion, both the registration date and mileage of a used car have a significant impact on its price in the market. A newer car with lower mileage is typically valued higher, while an older car with higher mileage is priced lower. When shopping for a used car, it is important to consider these factors in order to make an informed decision and get the best value for your money.

7. REFERENCES

RDocumentation, https://www.rdocumentation.org/. Accessed 10 February 2023.

Cross Validated, https://stats.stackexchange.com/. Accessed 10 February 2023.

Cohen, Yosef, and Jeremiah Y. Cohen. *Statistics and Data with R: An Applied Approach Through Examples*. Wiley, 2008.

"Factors Affecting Car Prices Research Paper." *WOW Essays*, 15 February 2020, https://www.wowessays.com/free-samples/research-paper-on-factors-affecting-car-prices/. Accessed 12 February 2023.

"4.6 - Normal Probability Plot of Residuals | STAT 462." *STAT ONLINE*, https://online.stat.psu.edu/stat462/node/122/. Accessed 10 February 2023.

Haan, Marco, and Peter Kooreman. *Price Anomalies in the Used Car Market*. 2002.

Heumann, Christian, and Michael Schomaker. *Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R*. Springer International Publishing, 2017.

Rice, John A. *Mathematical Statistics and Data Analysis*. Thomson/Brooks/Cole, 2007.