Average Price of a Car

Nico Hillison

Introduction

I will be researching and analyzing data from a file that contains information about 110 new car models made in the year 2020. What I will be trying to accomplish is creating two models to predict the average price of a car. One model will be more complex than the other. After creating both models, I will discuss which model is more useful and generally better. Other topics that will be discussed are the association between fuel economy and price for the new car models and the drive type of each car and price.

Data cleaning/variable information

Both of the models that had to be created were worked around the average price. To get the average price, I created a new variable named "avgPrice". I created this variable by adding together two columns (LowPrice and HighPrice) and then dividing it by two. When making a histogram of "avgPrice", I noticed that the histogram was not bell-shaped nor centered (skewed to the right). To fix this issue, I decided to take the log of "avgPrice" and make a new variable called "Log.avgP". In the histogram of "Log.avgP", the data seemed more centered and bell-shaped, that is why that is the variable used in both models which correctly predicts the average price. For the complex model, several second-order terms were added. I learned that Acc060 and Weight interacting with themselves had significance in predicting the average price so both were squared separately in the complex model, as well as the interaction between Weight:HwyMPG and Acc060:Weight. I found that HwyMPG by itself is not important to the average price. The two variables "Make" and "Model" are not necessary for any of our research

and are not being used at all. To prepare the data for analysis, there was no need to do any cleaning with the data. There were no missing data variables in any of the rows or columns.

Methods

To create both models (simple and complex), the statistical procedures that will be used will be linear models. The simple model was used to predict the average price (Log.avgP) using Seating, Drive, Acceleration 0-60, and Weight. The complex model was used to predict the average price (Log.avgP) using Seating, Drive, Acceleration 0-60, Weight, Highway Miles per Galon, Acceleration 0-60 squared, Weight squared, Weight:Acceleration 0-60, and Weight:Highway Miles per Galon. As well as creating and comparing both models with an ANOVA test, I am researching the possible association between fuel economy and price as well as drive type and price. For those two questions, I will conduct a 2-sample t-test to check on the correlation between fuel economy and price and I will conduct a Nested F-test to find if there is a correlation between drive type and price. All analyses were performed using the software package used (R) and [any add-on packages].

Results

The most important variable that was gathered was the Log of the average price (dependent variable). With that variable, we are able to check other variables which are independent such as: “HwyMPG”, “Seating”, “Drive”, “Acc060”, and “Weight”. The data table shown below provides us with the information on what variable makes the average price go either up or down depending on the variable used. The standard deviation for all of the variables is relatively the same.

General Results
Average Price

Variable	Mean	Maximum	Minimum	St. Dev
HwyMPG	-0.03839	1.0832	-0.7694	0.00477
Seating	0.0653	1.2565	-0.9653	0.0310
Drive (FWD&RWD)	-0.6022 & 0.1575	1.0271	-0.7273	0.0852 & 0.1715
Acc060	-0.2210	0.7970	-0.6910	0.0177
Weight	3.66e-04	1.391	-0.594	3.62e-05

Table 1. Summary of descriptive statistics for Average Price

For the simple model, I performed a linear model to explore the correlation between the "Log of the average price" and the variables: "HwyMPG", "Seating", "Drive", "Acc060", and "Weight". In this linear model, there are no second-order terms and it is fairly straightforward. This model has an Adjusted R-square value of 0.789, meaning that there is a strong correlation between those 5 variables and the log of average price.

```
Call:
lm(formula = car.df$Log.avgP ~ HwyMPG + Seating + Drive + Acc060 +
    Weight)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4151 -0.1393 -0.0051  0.1165  0.7396

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.57e+00   3.72e-01    9.58  6.2e-16 ***
HwyMPG       6.35e-03   5.23e-03    1.21  0.22741
Seating     -1.02e-01   2.68e-02   -3.82  0.00023 ***
DriveFWD    -1.76e-01   5.93e-02   -2.97  0.00371 **
DriveRWD    -1.48e-01   1.16e-01   -1.27  0.20553
Acc060      -1.26e-01   1.79e-02   -7.04  2.3e-10 ***
Weight       3.79e-04   5.74e-05    6.61  1.8e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.207 on 103 degrees of freedom
Multiple R-squared:  0.801,     Adjusted R-squared:  0.789
F-statistic: 68.9 on 6 and 103 DF,  p-value: <2e-16
```

Simple Model. Linear model showing the correlation between average price and several variables.

For the complex model, I performed a linear model to explore the correlation between the "Log of the average price" and the variables: "Seating", "Drive", "Acc060", "HwyMPG", and "Weight" as well as several second-order terms such as "Weight^2", "Acc060^2", "Acc060:Weight" and "Weight:HwyMPG". This model has an Adjusted R-square value of 0.833, meaning that there is also a strong correlation between those 9 variables and the log of average price.

```
Call:
lm(formula = car.df$Log.avgP ~ Drive + Weight + Acc060 + HwyMPG +
    Seating + I(Weight^2) + I(Acc060^2) + Acc060:Weight + Weight:HwyMPG)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3698 -0.1291 -0.0137  0.1076  0.5789

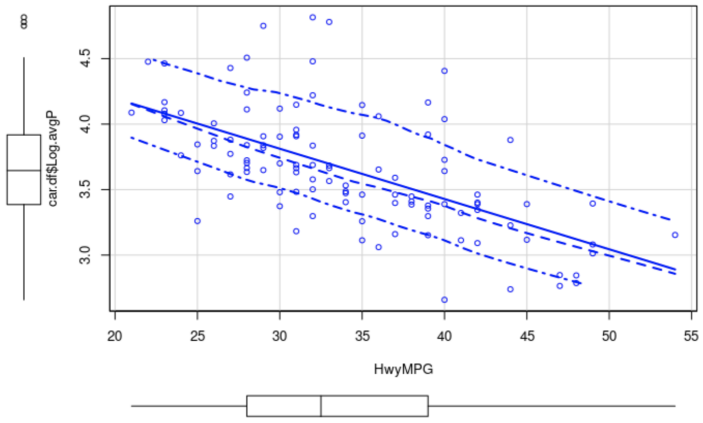
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.96e+00   1.72e+00    5.80  7.8e-08 ***
DriveFWD      -1.56e-01   5.47e-02   -2.84   0.0054 **
DriveRWD      -2.98e-01   1.20e-01   -2.49   0.0145 *
Weight        -1.28e-03   5.99e-04   -2.13   0.0357 *
Acc060        -8.33e-01   1.96e-01   -4.26  4.7e-05 ***
HwyMPG        -5.12e-02   2.65e-02   -1.93   0.0562 .
Seating       -8.17e-02   2.48e-02   -3.29   0.0014 **
I(Weight^2)    1.03e-07   4.46e-08    2.32   0.0222 *
I(Acc060^2)    3.48e-02   8.00e-03    4.35  3.4e-05 ***
Weight:Acc060  4.21e-05   2.42e-05    1.74   0.0846 .
Weight:HwyMPG  1.82e-05   7.31e-06    2.49   0.0144 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.184 on 99 degrees of freedom
Multiple R-squared:  0.848,     Adjusted R-squared:  0.833
F-statistic: 55.4 on 10 and 99 DF,  p-value: <2e-16
```

Complex Model. Linear model showing the correlation between average price and several variables as well as second-order terms.

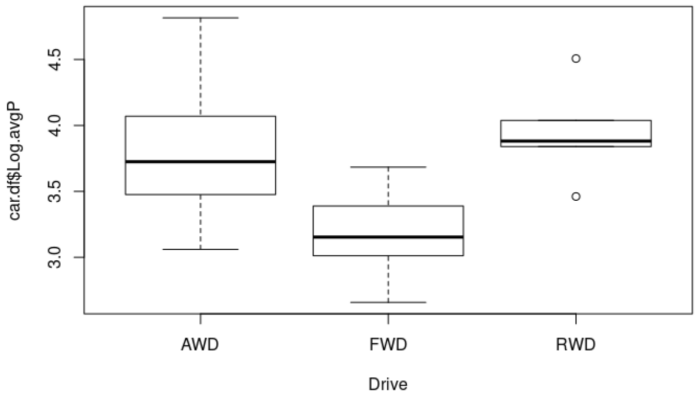
To see which model is better, I conducted an analysis of variance (ANOVA) test. The test statistics revealed that the complex model is better than the simpler model (p-value = 0.000017)

We also performed a 2-sided t-test to determine whether there is an association between fuel economy and price. The conditions for a 2-sided t-test were not met. The test revealed significant evidence that there is a correlation between fuel economy and price (p-value = 0.0000000000011).



Scatterplot. Correlation between fuel economy and price.

Finally, we performed a Nested F-test to determine whether there is an association between drive type and price. The conditions were met, so we used the Nested F-test statistics. The test revealed evidence that there is variability in drive type and price (p-value = 0.00000000044).



Boxplot. Between drive type and price

Discussion/Conclusion

The objective was to create two models (simple and complex) to find the average price of a car. Also, we had to figure out if there is any association between fuel economy and price as well as drive type and price. For the created models, by just looking at the Adjusted R^2 values of both models, I first thought that the difference between both was so minimal that the simpler model was better. But, after conducting an ANOVA test to compare both models to each other, I finally concluded that the complex model is better than the simpler model. For the correlation questions, through a 2-sample t-test, we were able to conclude that there is an association between fuel economy and price. Finally, after performing a nested F-test, we found that there is also an association between drive type and price. For future research, I would like to see how other variables such as "material" (what fabric is used in the car) or technology would have an effect on price. A limitation of this dataset was the small number of observations that the sample has (110). If we are generalizing this sample to the population of cars, I do not believe it would be very accurate. To better future research, we would increase the sample size to have a more accurate model for predicting price. In conclusion, the complex model is better than the simpler model to find the average price of a car and I found that there is a correlation between fuel economy and price as well as drive type and price.

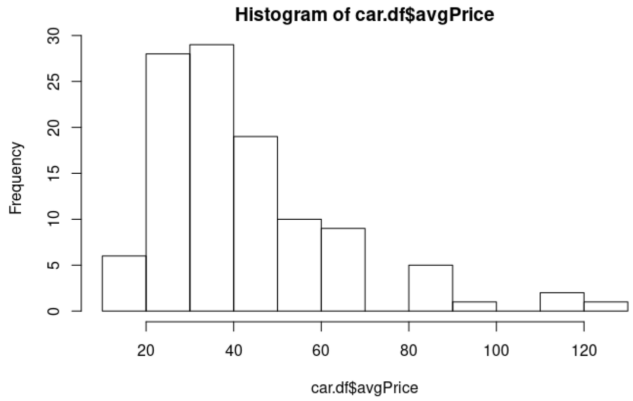
Appendix

<u>How I got the average price</u>

```
```{r}
car.df$avgPrice <- ((car.df$LowPrice + car.df$HighPrice) / 2)
car.df$Log.avgP <- log(car.df$avgPrice)
```

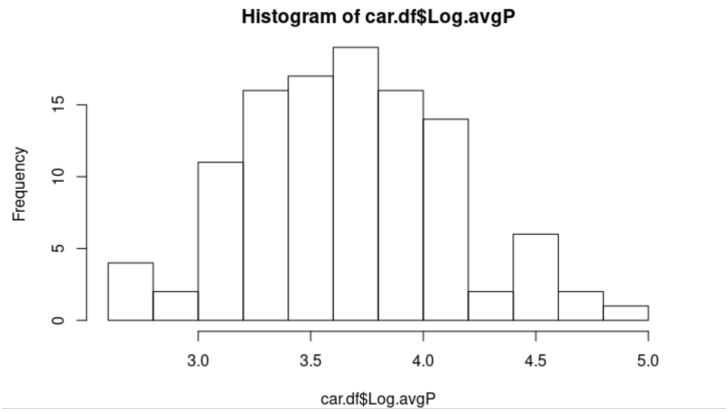
Creating the variable for the average price and taking its log
```

To find the average price I created a new variable named “avgPrice” and it consisted of the addition of the variables “LowPrice” and “HighPrice” divided by 2. When looking at the histogram of this new variable, I could tell that it was not bell-shaped and was skewed to the right.



Histogram. AvgPrice variable

Because the histogram did not meet the specifications needed, I decided to take the log of “AvgPrice” and make it into a new variable called “Log.avgP”. After making that change and creating a histogram of “Log.avgP” the shape was more bell-shaped and centered.



Histogram. Log.avgP variable

ANOVA test to compare both models (simple and complex):

1. Parameters: $\mu_3 \ldots \mu_{10}$ = both models
2. Hypothesis:
    a. H0 : $\mu_3 = \mu_4 \ldots = \mu_{10}$
    b. Ha : some $\mu_i \; != \mu_j$

2.1 Check Conditions:

- ➢ SRS (Simple Random Sample)
- ➢ Groups are independent
- ➢ Categories do have more than 30 samples
- ➢ The largest standard deviation is no more than 2 times the smallest.

Since the conditions are not met, we use an ANOVA statistic.

```
Analysis of Variance Table

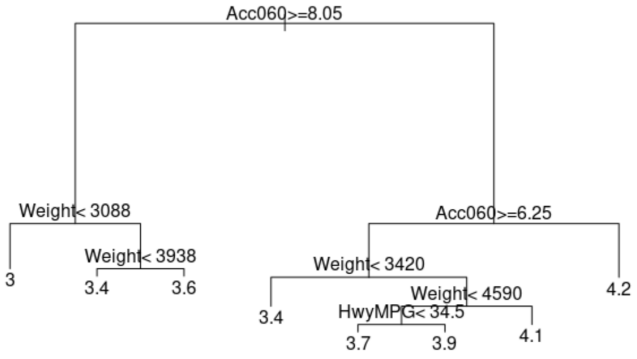
Model 1: car.df$Log.avgP ~ HwyMPG + Seating + Drive + Acc060 + Weight
Model 2: car.df$Log.avgP ~ Drive + Weight + Acc060 + HwyMPG + Seating +
    I(Weight^2) + I(Acc060^2) + Acc060:Weight + Weight:HwyMPG
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1    103 4.42
2     99 3.36  4      1.06 7.79 1.7e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA table. Comparing simple vs. complex models.

2.2 F = 7.79

3. p-value = 0.000017

4. We reject the null hypothesis because the p-value is so small (< 0.05) and conclude that the complex model is better than the simpler model.



Regression Tree. Find connections between variables

The regression tree showed that there are interactions between the variables Acc060 and Weight (Acc060:Weight in the linear model) as well as Weight and Highway Miles per Gallon (Weight:HwyMPG in the linear model).

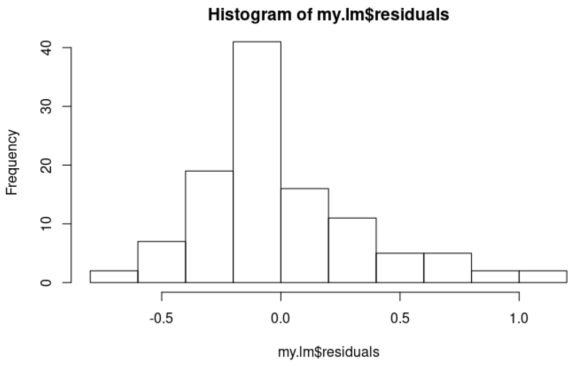
<u>Test for correlation between fuel economy and price:</u>

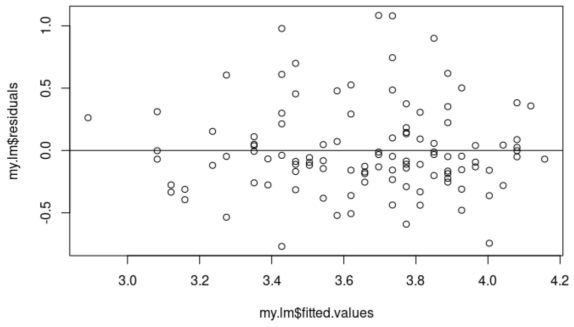
2. Parameters: x = HwyMPG, y = Log.avgP

3. Hypothesis:
    i. Ho : ρ = 0
    ii. Ha : ρ =/ 0

2.1. Check conditions: LINES

- ➔ Linearity: test can be performed if the relationship is approximately linear. A linear relationship is the best approximation for the two variables.
- ➔ Independence: The y values in our data are independent of each other.
- ➔ Normality: The residuals are not normally distributed, and the distribution is slightly skewed to the right, so LINES fails (see the distribution of residuals).
- ➔ Equal Variance: The residuals of the linear regression are not fanned or curved, but are not evenly distributed above and below the line (RVF plot).
- ➔ S (Randomization): The data was collected using an SRS.





The LINES rules are not met, so we can use a t-test statistic and must use a linear model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.96331    0.16557   29.98  < 2e-16 ***
HwyMPG      -0.03839    0.00477   -8.06  1.1e-12 ***
```

Linear Model Coefficients. Correlation between fuel economy and price.

- T-value = -8.06
- P-value = 0.0000000000011

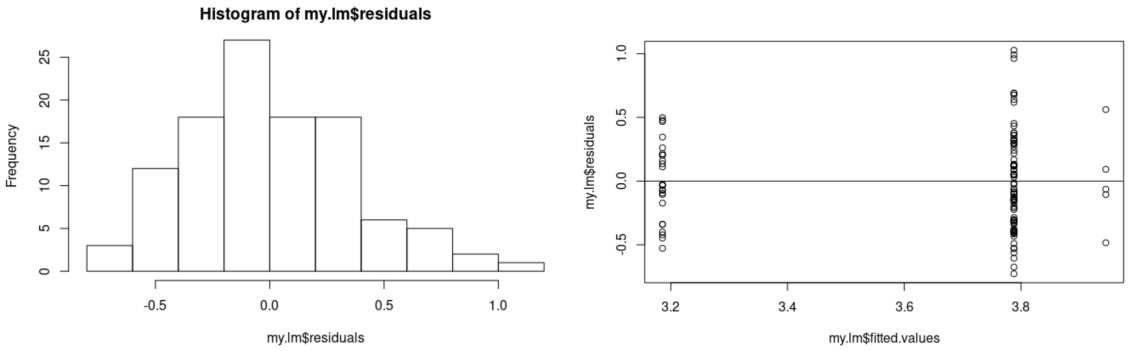
Conclusion: Because the p-value is so small (< 0.05) we reject the null hypothesis and conclude there is a strong association between fuel economy and price.

<u>Nested F-test for drive type and price:</u>

1. Parameters: x = Drive, y = Log.avgP
2. Hypothesis:
    b. Ho : $\mu_3 = \mu_4 \ldots = 0$
    c. Ha : some μ_i != 0

2.1. Check conditions: LINES

- ➔ Linearity: test can be performed if the relationship is approximately linear. A linear relationship is the best approximation for the two variables.
- ➔ Independence: The y values in our data are independent of each other.
- ➔ Normality: The residuals are normally distributed, but the distribution is slightly skewed to the right, so LINES fails (see the distribution of residuals).
- ➔ Equal Variance: The residuals of the linear regression are not fanned or curved and are evenly distributed above and below the line (RVF plot).
- ➔ S (Randomization): The data was collected using an SRS.



Since the conditions are not met, we use a Nested F-test statistic.

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Drive        2   7.34    3.67    26.5 4.4e-10 ***
Residuals  107  14.80    0.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nested F-test. Correlation between drive type and price.

2.2 F = 26.5

3. p-value = 0.00000000044

4. We reject the null hypothesis because the p-value is so small (< 0.05) and conclude that there is a strong correlation between drive type and price.