Philip Booth and Lora LaRochelle

Prof. Ning

SC321

November 18th, 2024

Data Analysis Report

Introduction:

The price of a car is dependent on a number of factors, including the manufacturer and the model of the car. Various other qualities inherent to the vehicle may also influence its price, such as its fuel economy, rate of acceleration, seating capacity, and more. In this paper, we seek to create a model that incorporates these various factors to predict the average price of a vehicle across all makes and models. The data used to conduct this analysis contain information on 110 new car models that were produced in 2020. The objective of this paper is to construct two models using vehicle qualities, one being fuel economy, to effectively predict the average price of a car, and then determine which model is better for our purposes.

Methods:

The key variables in this dataset are described in Table 1. We created a variable called “AvgPrice” which is the average of the low and high price variables. This was the response variable in our model.

Table 1: Key variables in the dataset and their descriptions.

Variable	Description
Make	The name of the manufacturing company of the vehicle.
Model	The specific name given to the vehicle by the manufacturing company.
LowPrice	The lowest price that the car sells for, in thousands of dollars.
HighPrice	The highest price that the car sells for, in thousands of dollars.
HwyMPG	The highway miles-per-gallon of the vehicle.
Seating	The seating capacity of the vehicle.
Drive	The type of drive of the car: either all-wheel, front-wheel, or rear-wheel.

Acc060	The time in seconds it takes for the vehicle to go from 0 to 60 miles-per-hour.
Weight	The weight of the vehicle, in pounds.
AvgPrice	The average of the low price and the high price of the vehicle, also in thousands of dollars: AvgPrice = (LowPrice + HighPrice)/2.

We were most interested in investigating the association between the average price of the vehicle and its fuel economy, which is the variable HwyMPG. We also wanted to add another variable to our model to improve its prediction capacity. We selected this variable by looking at a correlation matrix of the quantitative variables in the dataset and selecting the variable most correlated with average price, which was the variable Acc060. A log-transformation to the average price variable was needed to satisfy the conditions for multiple linear regression. We began by creating a simple multiple linear regression model including fuel economy and Acc060 to predict the log of average price. We then constructed a full second-order model that included both variables, their quadratic terms, and an interaction term as predictors for the log of average price. An ANOVA nested-F test was used to compare these models and determine which model was better for our purposes.

Results (Part 1):

Descriptive statistics for the quantitative variables in the dataset are provided in Table 2. For the categorical variable Drive, 80 vehicles (72.72%) were all-wheel drive, 25 (22.72%) were front-wheel drive, and 5 (4.54%) were rear-wheel drive. A histogram of the average vehicle price shows that this variable is skewed to the right (Figure 3), and the log-transformed average price is shown in Figure 4. The correlation between the log of average price and fuel economy (HwyMPG) is -0.613 and the correlation between the amount of seconds needed to accelerate from zero to 60 miles per hour (Acc060) and the log of average price is -0.769, so both variables are negatively correlated with the log of average price. There was multicollinearity detected between Weight and HwyMPG (r=-0.817, VIF=6.06). Scatter plots showing the relationship between fuel economy and acceleration rate against the log of the vehicle's average price are included as Figures 5 and 6, respectively. The scatter plots show a linear relationship that could potentially be better approximated by a quadratic model, which leads us to believe that a full

second-order model may be necessary to adequately predict average price from these two variables in a multiple linear regression model.

Table 2: Descriptive statistics for the key quantitative variables in the dataset.

	Minimum	Q1	Median	Q3	Maximum	IQR	Mean	SD
HwyMPG	21.00	28.25	32.5	39.00	54.00	10.75	34.00	7.193
Acc060	4.100	6.80	7.70	8.675	12.10	1.875	7.835	1.568
AvgPrice	14.28	29.57	38.29	50.23	123.35	20.66	43.05	21.36
Weight	2085	3292	3845	4476	6100	1184	3875	21.36
Seating	2	5	5	6.5	9	1.5	5.527	1.37

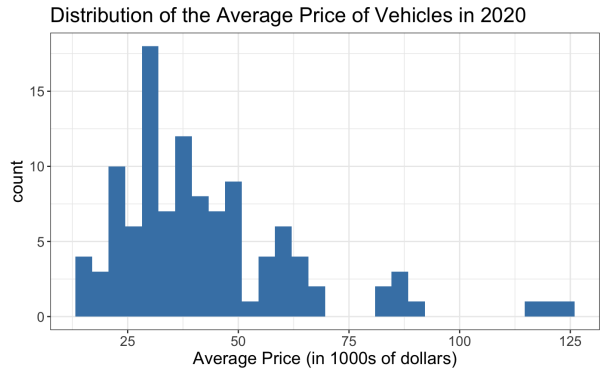


Figure 3: Histogram of the variable AvgPrice, which has a rightward skew. This plot shows the need for a log-transformation of this variable in the multiple linear regression model.

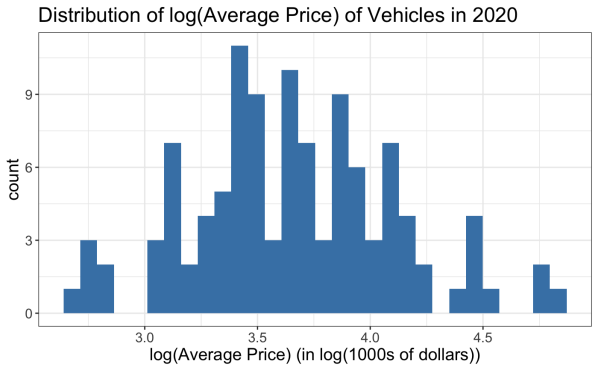


Figure 4: Histogram of the log-transformed AvgPrice, which appears to be bell-shaped and roughly symmetric.

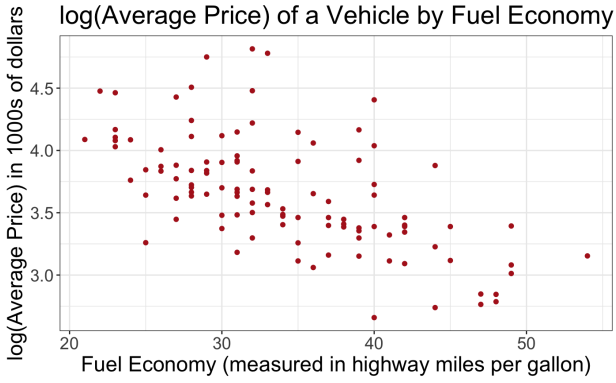


Figure 5: Scatter plot depicting the relationship between log(AvgPrice) and HwyMPG.

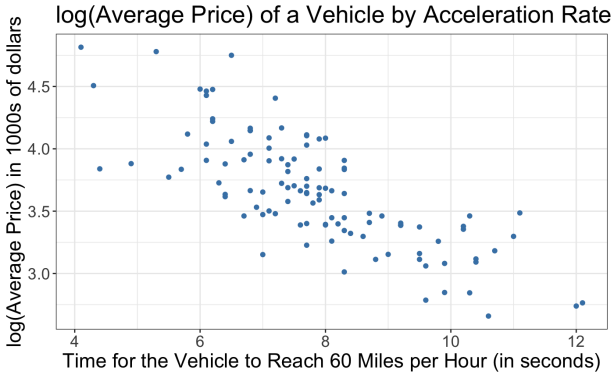


Figure 6: Scatter plot depicting the relationship between log(AvgPrice) and Acc060.

Results (Part 2):

The simple model predicting the log of average price from acceleration rate and fuel economy is given as $\widehat{log(AvgPrice)} = 5.764 - 0.21(HwyMPG) - 0.178(Acc060)$. Both the coefficients for $HwyMPG$ and $Acc060$ are significant with respective p-values 3.01e-07 and <2e-16. There is no detected multicollinearity between the predictors (VIF = 1.25), so the individual t-tests/p-values can be trusted. The R^2 for this model is 0.6803. The full second-order model, which includes both quadratic terms and the interaction term between our quantitative variables, is given as

$\widehat{log(AvgPrice)} = 5.82 - 0.007(HwyMPG) - 0.26(Acc060) + 0.00054(HwyMpg)^2$

$+ 0.0194(Acc060)^2 - 0.0065(HwyMPG * Acc060)$. The individual t-tests show that $Acc060$ and $Acc060^2$ are significant in the model with respective p-values 0.0277 and 0.0416. After the second-order terms are added, the coefficient for $HwyMPG$ is no longer significant (p-value=0.8182). The R^2 for this model is 0.6936 and the adjusted R^2 is 0.6789, showing a very slight improvement from the adjusted R^2 of the simple model, which was 0.6743. A nested-F test comparing the full model against the second model showed that the full model was not significantly more effective at predicting average price (F=1.5103, p-value=0.2162). We conclude that the better model is our original, simple model with no second-order terms to align with the statistical principle of parsimony.

Discussion:

We found that 68.03% of the variability in the log of average price can be explained by the vehicle's fuel economy and acceleration rate and 69.36% of the variability can be explained by the full second-order model. We prefer the simpler model as it was found to be adequate in predicting log(average price), with the prediction equation $\widehat{log(AvgPrice)} = 5.764 - 0.21(HwyMPG) - 0.178(Acc060)$. Both fuel economy and acceleration rate are negatively associated with log(average price), suggesting that

the better the fuel economy and higher the acceleration rate of the vehicle, the lower its log of average price will be.

While one may think that vehicles with high values in these variables would be more expensive, our results suggest the opposite. This is good for consumers as it indicates that choosing a more fuel-efficient vehicle could save them money in both the short-term and long-term. We recommend that car-buyers looking to save money should consider purchasing a vehicle with a high fuel economy and rate of acceleration from zero to 60 miles per hour.

One possible limitation of this analysis is that all of our data came from 2020. We therefore cannot generalize our results to more recent years. A potential avenue of future research could be comparing the relationship between the variables in the model across multiple years: do fuel economy and the amount of time the car takes to accelerate to 60 miles per hour do a better job of predicting the average price of a vehicle in different years? Are there years where other quantitative predictors are more highly correlated with its average price?

Based on this 2020 data, we can conclude that there is a significant negative relationship between the average price of a vehicle and a linear combination of its fuel economy and the time in seconds that the vehicle takes to accelerate from zero to sixty seconds in this year. A simple model with no second order terms is adequate at describing this relationship, and our results could be helpful to potential car-buyers.

Appendix:

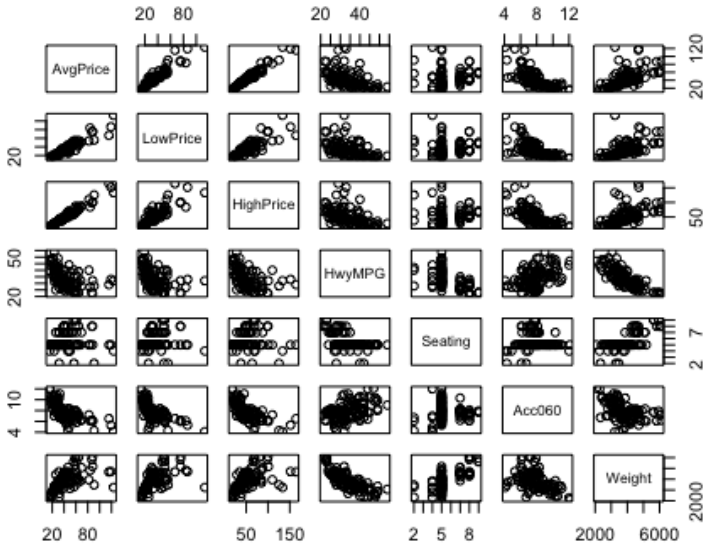
To determine which predictors to include in our first, simple model, we consider which variables are most correlated with average price. It appears that $Acc060$ is most correlated with $AvgPrice$ with a correlation of $r = -0.696$. Since the researchers are most interested in understanding the relationship between fuel economy (the variable $HwyMPG$ in the dataset) and $AvgPrice$, we construct a simple model predicting $AvgPrice$ from the predictors $HwyMPG$ and $Acc060$.

There is detected multicollinearity between Weight and HwyMPG (VIF=6.06), so Weight was not included in our model.

```
m1 <- lm(AvgPrice ~ HwyMPG + Seating + Acc060 + Weight, data=dat)
vif(m1)

##   HwyMPG  Seating   Acc060   Weight
## 3.116637 3.020586 1.826353 6.063015

numcols <- dat[sapply(dat, is.numeric)]
pairs(AvgPrice ~. , data=numcols)
```

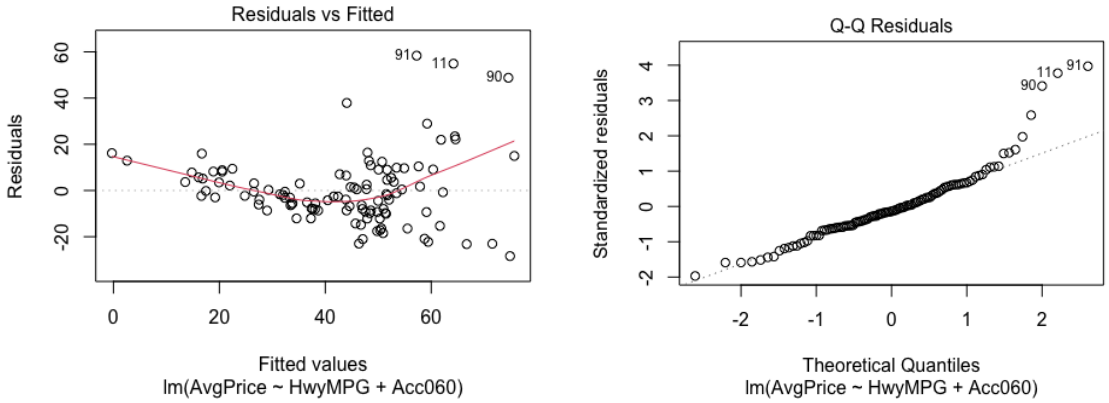



```
cor(numcols[3:7])

##              HwyMPG     Seating      Acc060     Weight   AvgPrice
## HwyMPG    1.0000000 -0.52119935  0.44968160 -0.8173182 -0.4972649
## Seating  -0.5211994  1.00000000  0.01831626  0.7192529  0.1172199
## Acc060    0.4496816  0.01831626  1.00000000 -0.4522753 -0.6962995
## Weight   -0.8173182  0.71925287 -0.45227526  1.0000000  0.5875611
## AvgPrice -0.4972649  0.11721990 -0.69629954  0.5875611  1.0000000
```

In trying to predict AvgPrice from HwyMPG and Acc060, the conditions to run a multiple linear regression do not check. Normality and constant variance of the residuals are not satisfied. There is a clear fanning-out pattern in the residual vs. fitted plot.

```
bad_model <- lm(AvgPrice ~ HwyMPG + Acc060, data=dat)
plot(bad_model)
```



We thus log-transform the variable $AvgPrice$ to and check the conditions for the multiple regression model $log(AvgPrice) = \beta_0 + \beta_1(HwyMPG) + \beta_2(Acc060) + \epsilon$, which yields the predicted equation $\hat{log(AvgPrice)} = 5.764 - 0.21(HwyMPG) - 0.178(Acc060)$.
Linearity: The residuals vs. fitted plot looks randomly scattered after the log transformation is applied.

Zero mean: Since we used the least-squares regression methods, the residuals will have a mean of 0.

Constant variance of residuals: the variance of residuals appears to be approximately constant.

Normality of residuals: The normal quantile plot looks linear, with a few possible outliers showing slight deviations from the overall general linear trend.

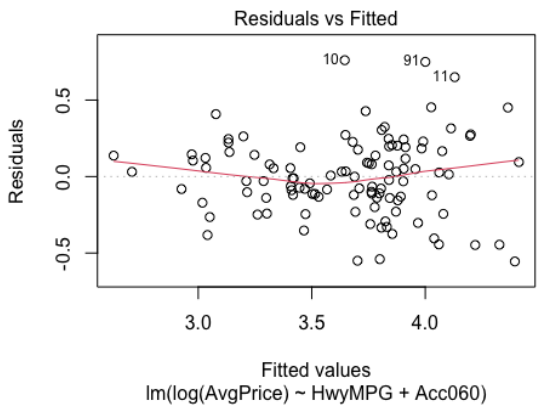
Independence & Randomness: these conditions are required for formal inference. We have no reason to believe that the data do not come from a random sample, and we will conduct this analysis under the assumption that the data are independent from one another.

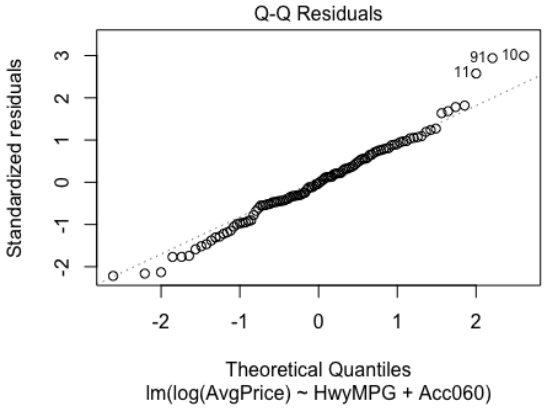
In our simple model $\hat{log(AvgPrice)} = 5.764 - 0.21(HwyMPG) - 0.178(Acc060)$, both the coefficients for $HwyMPG$ and $Acc060$ are significant with respective p-values 3.01e-07 and <2e-16. There is no detected multicollinearity between the predictors (VIF = 1.25), so the individual t-tests/p-values can be trusted. The R^2 for this model is 0.6803.

```
simple.cars.lr <- lm(log(AvgPrice) ~ HwyMPG + Acc060, data=dat)
summary(simple.cars.lr)
```

```
##
## Call:
## lm(formula = log(AvgPrice) ~ HwyMPG + Acc060, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55507 -0.13671 -0.00599  0.16460  0.76090
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.763699   0.142957  40.318  < 2e-16 ***
## HwyMPG      -0.020965   0.003834  -5.468 3.01e-07 ***
## Acc060      -0.177760   0.017595 -10.103  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2572 on 107 degrees of freedom
## Multiple R-squared:  0.6803, Adjusted R-squared:  0.6743
## F-statistic: 113.8 on 2 and 107 DF,  p-value: < 2.2e-16
```

```
plot(simple.cars.lr)
```






```
vif(simple.cars.lr)
```

```
##   HwyMPG   Acc060
## 1.253468 1.253468
```

We next construct a more complex model, adding in both second order terms and the interaction terms. We begin by assessing the conditions for this more complex, full second-order model:

Linearity: The residual vs. fitted plot shows a clear linear trend.

Zero mean: Since we used the least-squares regression method, the residuals will have a mean of 0.

Constant variance: The residual vs. fitted plot shows consistent spread.

Normality of residuals: The normal Q-Q plot looks generally linear in form, with a few possible outliers.

Independence & Randomness: we conduct this analysis under the assumption that these requirements are satisfied. The full second-order model is

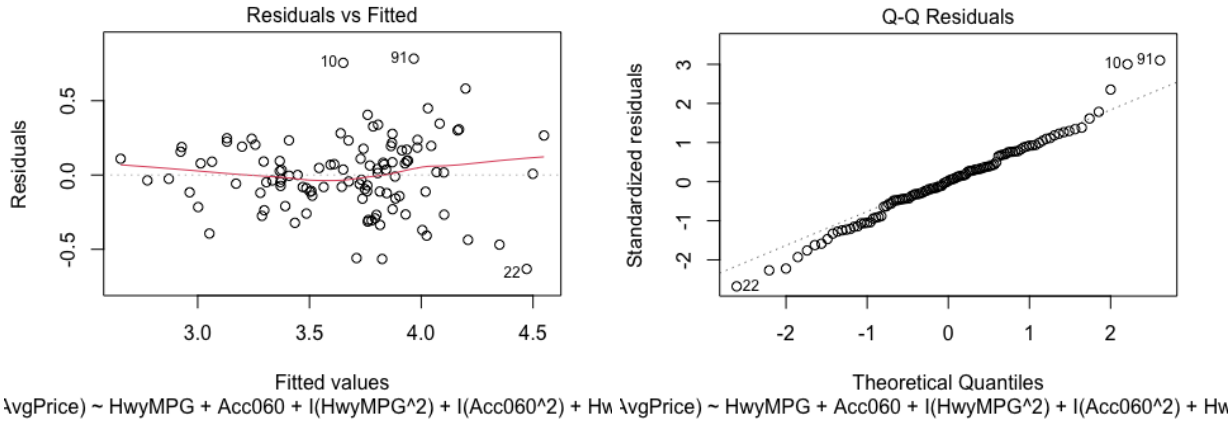
$$\hat{log(AvgPrice)} = 5.82 - 0.007(HwyMPG) - 0.26(Acc060) + 0.00054(HwyMpg)^2 + 0.0194(Acc060)^2 - 0.0065(HwyMPG * Acc060).$$

The individual t-tests show that $Acc060$ and $Acc060^2$ are significant in the model with respective p-values 0.0277 and 0.0416. After the second-order terms are added, the coefficient for $HwyMPG$ is no longer significant (p-value=0.8182). The R^2 for this model is 0.6936 and the adjusted R^2 is 0.6789, showing only a very slight improvement from the adjusted R^2 of the simple model, which was 0.6743.

```
complex.cars.lr <- lm(log(AvgPrice) ~ HwyMPG + Acc060 + I(HwyMPG^2) +
I(Acc060^2) + HwyMPG*Acc060, data=dat)
summary(complex.cars.lr)

##
## Call:
## lm(formula = log(AvgPrice) ~ HwyMPG + Acc060 + I(HwyMPG^2) +
##     I(Acc060^2) + HwyMPG * Acc060, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63231 -0.12179  0.00404  0.16906  0.78336
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.8202914  0.7471773   7.790 5.28e-12 ***
## HwyMPG        -0.0070676  0.0306763  -0.230   0.8182
## Acc060        -0.2596481  0.1162470  -2.234   0.0277 *
## I(HwyMPG^2)    0.0005404  0.0004941   1.094   0.2766
## I(Acc060^2)    0.0194358  0.0094230   2.063   0.0416 *
## HwyMPG:Acc060 -0.0065597  0.0036907  -1.777   0.0784 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2554 on 104 degrees of freedom
## Multiple R-squared:  0.6936, Adjusted R-squared:  0.6789
## F-statistic:  47.1 on 5 and 104 DF,  p-value: < 2.2e-16

plot(complex.cars.lr)
```



To compare these models, we will conduct a nested-F test pairing the full second-order model against the simple model with no second order terms.

Let $\beta_3, \beta_4, \beta_5$ be the model coefficients for $HwyMPG^2$, $Acc060^2$, and $HwyMPG * Acc060$, respectively.

$H_0: \beta_3 = \beta_4 = \beta_5 = 0.$

H_a: at least one $\beta_i \neq 0$ for $i = 3, 4, 5$.

A nested-F test gives us an F-statistic of 1.5103 and subsequent p-value of 0.2162. We thus fail to reject the null hypothesis at a 0.05 significance level. We do not have enough evidence to suggest that the full second-order model is more effective than the simple model. We thus prefer the simple model to align with the statistical principle of parsimony.

```
anova(simple.cars.lr, complex.cars.lr)

## Analysis of Variance Table
##
## Model 1: log(AvgPrice) ~ HwyMPG + Acc060
## Model 2: log(AvgPrice) ~ HwyMPG + Acc060 + I(HwyMPG^2) + I(Acc060^2) +
##     HwyMPG * Acc060
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    107 7.0783
## 2    104 6.7828  3   0.29549 1.5103 0.2162
```