w203 Lab 2 Final Report

Do people actually care about horsepower when buying a car?

Aruna Bisht, Don Irwin, Kurt Eulau, Qin Luo

12/6/2021

Github repository: https://github.com/mids-w203/w203_regression_lab_bottlenose

1. Introduction

Car manufacturers spend close to $100 billion per year globally on the research and development of vehicles.[1] Features of cars that companies focus on can have huge impacts for companies and consumers alike. Horsepower has long been an often touted characteristic of a car in the United States and certainly some Americans consider horsepower when they purchase a new vehicle.

Figure 1: American Car Advertisements

The underlying relationship between horsepower and sales, however, remains unclear. To shed light on this topic, our research question is the following:

How did the horsepower of a car affect the sales of that model in the United States in 2017?

In the case that we find that horsepower seems to at least somewhat explain car sales, then manufacturers might consider investing further in increasing horsepower to increase sales. In the case that no strong link can be found, car makers could look at other features of their automobiles in an attempt to improve sales volume.

2. Data and Research Design

We utilize two data sets: one data set containing features of a manufacturers' vehicle by make and model and a second containing total sales by model for a given year.

Car Model Features by Year: This is publicly available on Kaggle and contains vehicle features by model, and trim line. We use the data for the year 2017, which is the latest year for which there is data in the dataset.

https://www.kaggle.com/CooperUnion/cardataset/version/1

Car Sales by Year: Sales volume data for each car model (though not trim line) was available at http://www.carsalesbase.com. Since the data did live in a unified data set, we scraped the data from the site and made it available at: https://www.kaggle.com /donirwin/car-sales-by-model-in-the-united-states-2017.

Data Manipulation, Mapping and Cleansing: *Car Model Features by Year* contains information by model and trim while *Car Sales by Year* contains information only by model. We filter *Car Model Features by Year* to include only records for a single year (2017) – this yielded a total of *1668* records. We standardize vehicle names so that the two data sets could be joined.

We collapse each model's trim records from *Car Model Features by Year* into a single model record, taking averages of the features horsepower, MSRP (manufacturer's suggested retail price), city, highway, and combined MPG (miles per gallon) while preserving features that are the same across all trim lines, such as vehicle size. Applying this collapsing logic reduced our record count from *1668* records to *231* records.

Our process yielded *231* records, of which, *184* were successfully joined, and constitue our dataset. 47 were unable to be joined and are not included in our data set.

Introduction of IS_LUXURY variable: We designated eleven (11) vehicle brands as luxury vehicles. This methodology is inherently flawed as it perhaps omits some vehicles which are in fact luxury vehicles, and includes some makes, of otherwise luxury brands, which are not luxury vehicles are such.

Hard Coding Size of 7 Vehicle Models: It was necessary for us to manually assign the "size" variable to 7 vehicles in our dataset.

The top five selling vehicle make and models from our combined dataset are shown below

Table 1: Top Five Selling Vehicles in USA, 2017

Make and Model	Sales	Avg HP	Avg Cmb MPG	Avg MSRP	Size	Trans Type	Is Luxury?
chevrolet silverado	1171728	313.0000	19.52000	39816.00	large	automatic	N
ford f series	896764	335.9545	20.03409	41609.61	large	automatic	N
chevrolet equinox	580916	182.0000	25.71429	27564.29	compact	automatic	N
toyota rav4	407594	180.1538	27.11538	30681.15	midsize	automatic	N
nissan rogue	403465	170.0000	29.00000	27015.00	midsize	automatic	N

The full list of variable names and definitions are list below for the sake of clarity and transparency:

Variable Names and Definitions:

Table 2: Variable Names and Definitions

variable_name	variable_definition
avg_hp	Average Horsepower: Also referred to as horsepower.
avg_hw_mpg	Average Highway Miles Per Gallon
avg_city_mpg	Average City Miles Per Gallon
size	Vehicle Size
avg_cmb_mpg	Average Combined Miles Per Gallon
trans_type	Transmission Type

Research Design

We will create a series of regression models within a causal theory in an attempt to explain vehicle sales. The goals for our models are listed below:

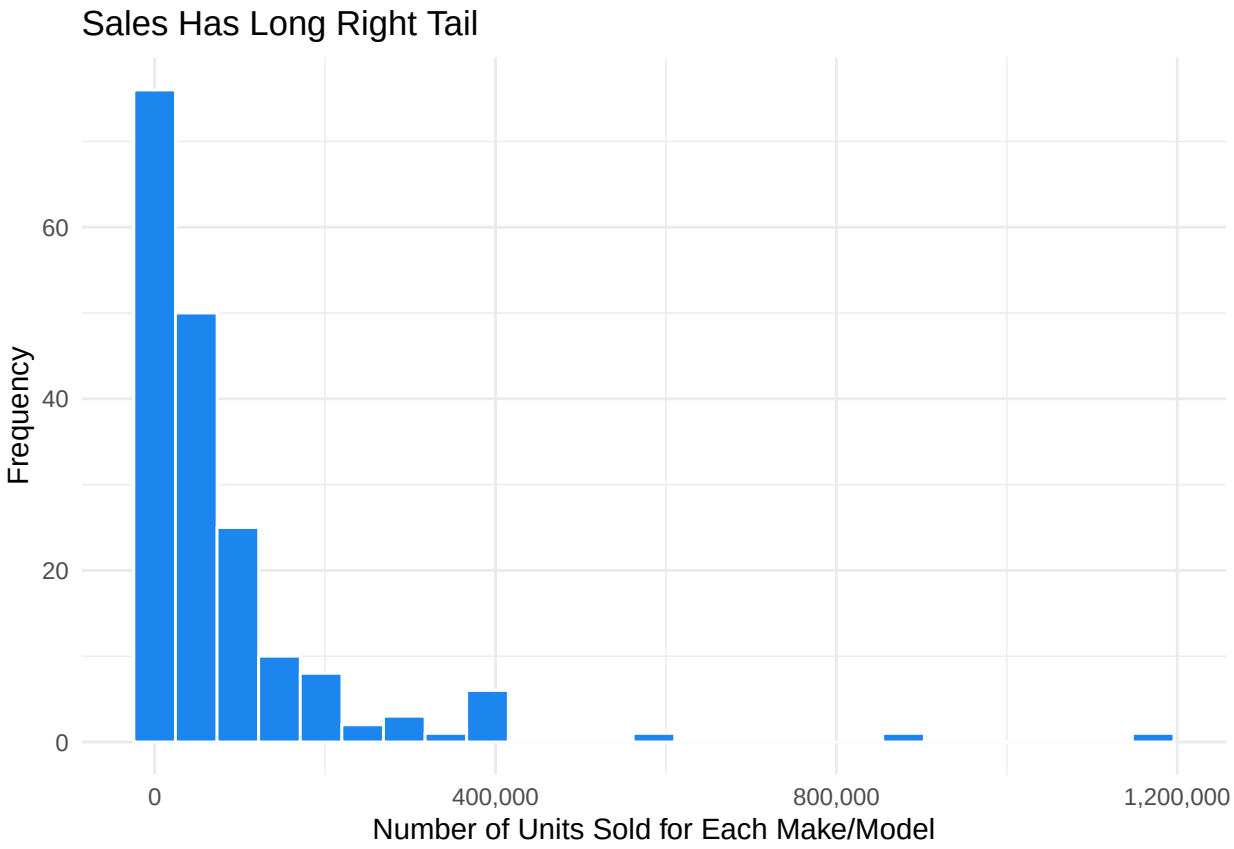
- Our primary concern is creating a model that can help state if horsepower causes sales to increase, decrease, or has no effect. Statistically, this means we need to find the statistical significance and effect size, if applicable, of horsepower as an

explanatory variable in a model to explain sales. When we say "cause" we explicitly acknowledge that we are moving from a perspective of associations to causal relationships.

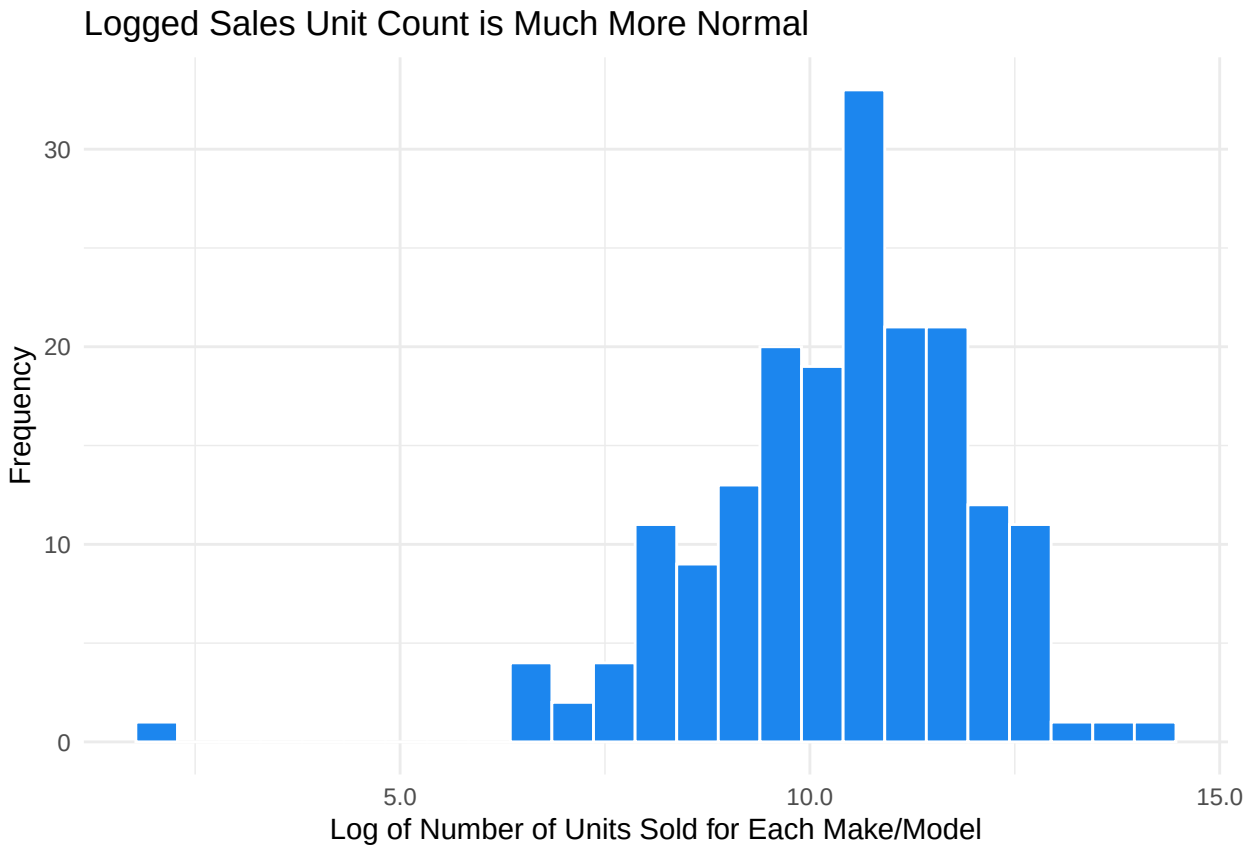
- We aim to create a model that explains enough variance to be useful.
- Defensibly and logically introduce variables other than horsepower to see if horsepower is "robust against the introduction of covariates." We pay special attention to the magnitude of the horsepower coefficient, where a stable coefficient as covariates are introduced indicates evidence that the effect of horsepower is "robust" and otherwise not "robust."
- As more variables enter the frame, we seek to articulate the causal relationships among variables to inform assertions regarding the causal relationships among variables.

2a Model Building Process

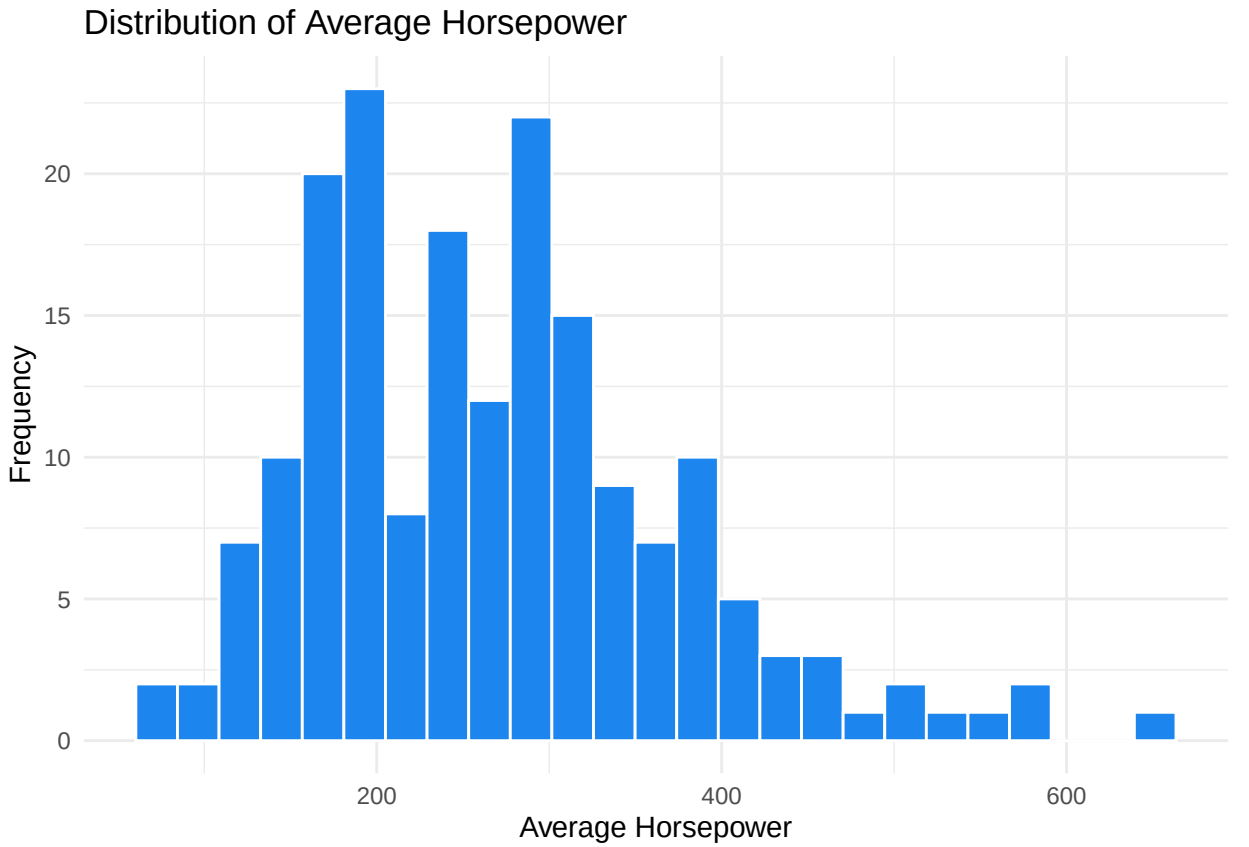
We begin with some EDA. We start with our response variable, number of cars sold, for each make/model combination, referred to as "sales".



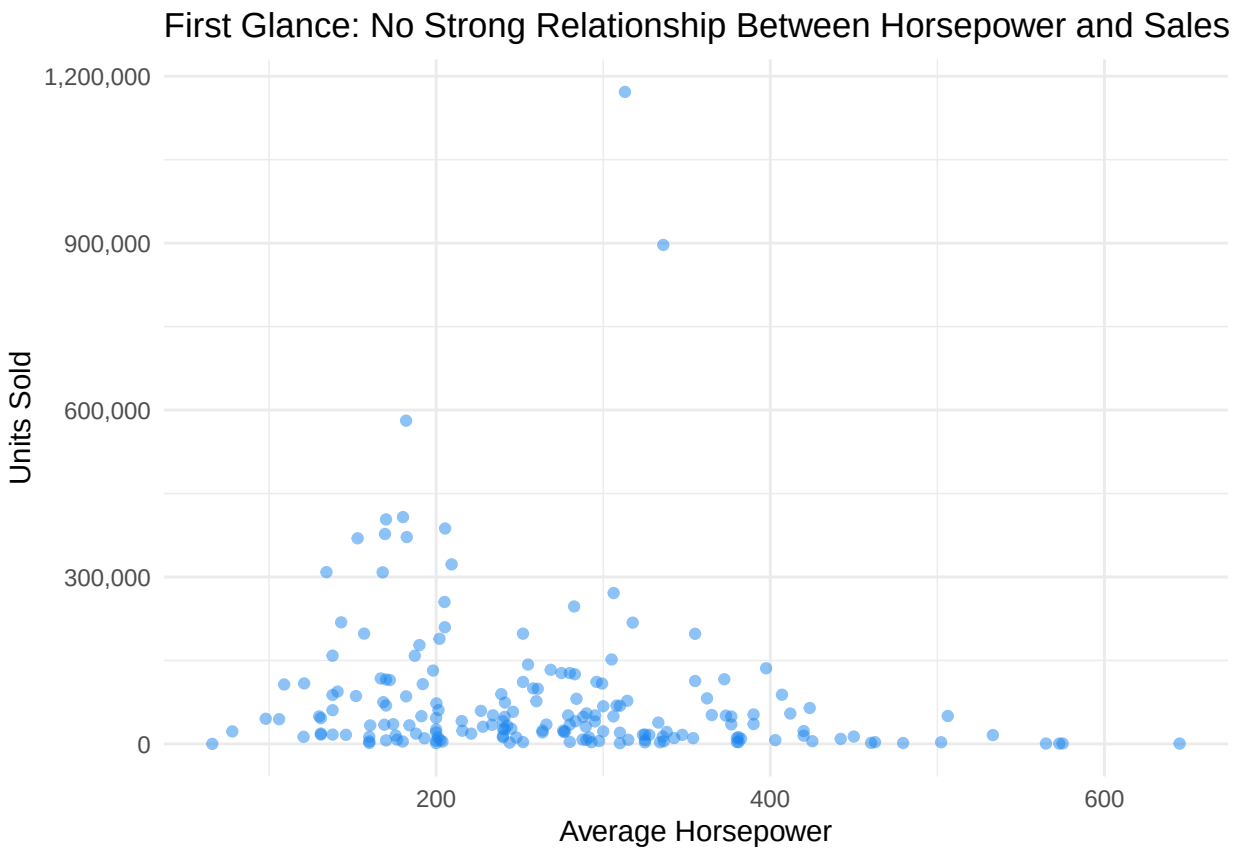
Sales has a long right tail. When using linear regression models to capture the relationship between sales and our RHS variables, we may consider a log transformation. Our sales data meets log transformation criteria (see Async Lesson 10.9) which we apply to improve model performance. Applying a natural log transformation distribution of logged sales unit are shown below:



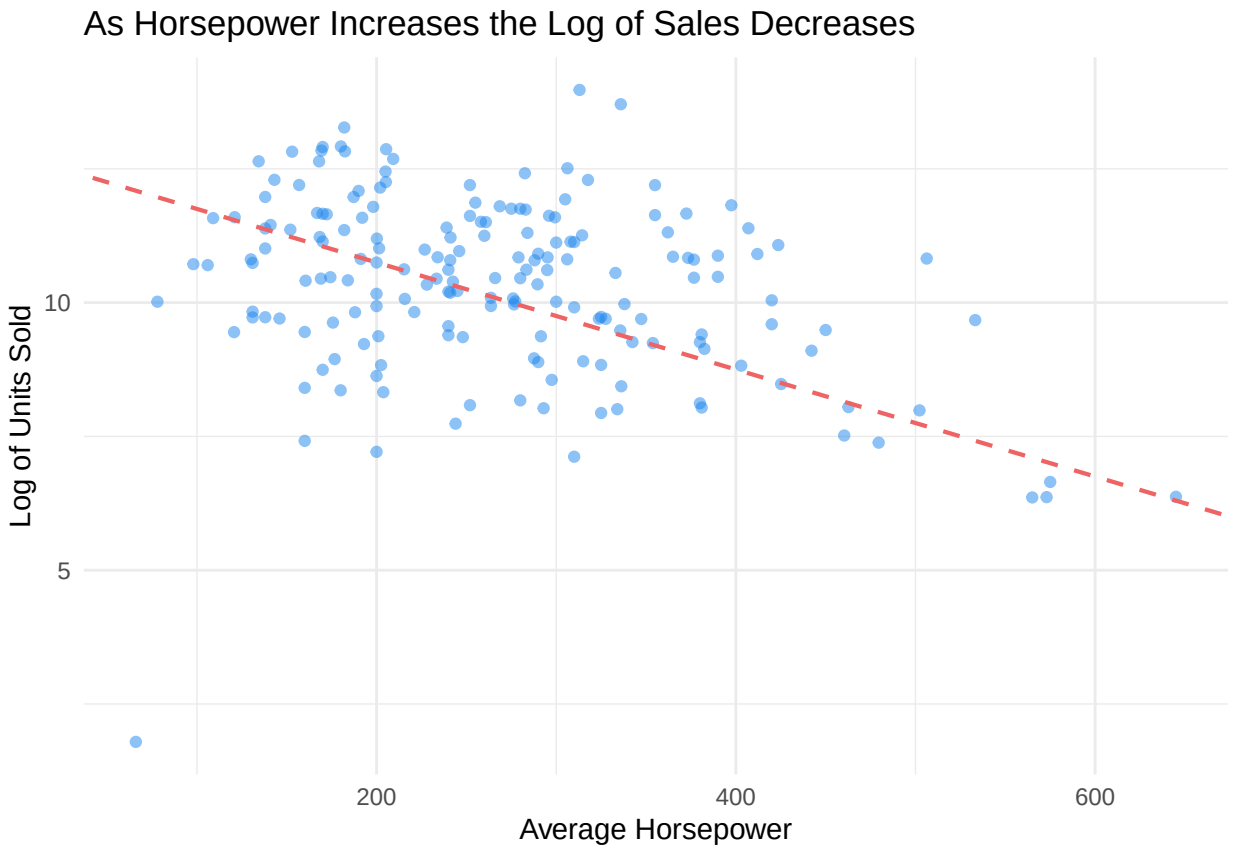
Now that we have examined sales we can turn to our principle explanatory variable, `avg_hp` and its distribution as shown below:



The distribution for `avg_hp` has a long right tail. The case for logging horsepower is less strong; the distribution scales fewer orders of magnitude and is not clustered around zero. Understanding both horsepower and sales, we can consider how they "move together." We turn our attention to a scatterplot of average horsepower and sales, as seen below:



There is little tendency for sales to move as horsepower does. Sales may increase as horsepower increase up to about 200 horsepower, at which point increases in horsepower might be associated with decreases in sales. The overall pattern is that most cars of differing levels of horsepower tend to sell less than 150,000 units. In fact, roughly 86% of the vehicles in the dataset (158 make/models) are sold at a volume less than 150,000 units. For the curious, the two vehicles with at least 900,000 sales are the Chevrolet Silverado (1,171,728 units sold) and the Ford F-Series Trucks (896,764 units sold). We look at the scatterplot below, of horsepower and log transformed sales to see if we can clarify any possible relationship.



Average horsepower and the log of sales "move" in a general pattern: as horsepower increases, the log of sales tends to decrease. A Toyota Camry, has average horsepower of 205 hp, a sales volume of 387,081 units (log sales of 12.86), and an average MSRP of $27,466. In contrast, the Dodge Viper has 645 horsepower, sales volume of only 585 (log sales of 6.37), and an average MSRP of $101,295. It may be that as horsepower increases, so does the cost of the car, which drives down sales[3], all else constant. This trend becomes increasingly clear if you ignore The Mitsubishi i-MiEV (horsepower of 66 hp and logged sales of 1.79), which had only 6 sales in the United States and was discontinued in the United States starting in 2018[2]. We will remove it from our data set before fitting our models.

Horsepower

Our first model, examines the relationship between the log of sales and average horsepower and takes the following form:

$$\text{MODEL 1} \rightarrow log(\text{sales}) = \hat{\beta}_0 + \hat{\beta}_1 \text{horsepower} + \hat{\epsilon}$$

Here, we find that both the intercept and coefficient for average horsepower to be statistically significant. We elect to use HC3 version of standard error calculations (the default used in `lmtest`'s `coeftest()` and `summary()`). Long and Ervin[4] recommend that (1) analysts correct for heteroeskedasticity whenever they suspect it, (2) that decision to correct should not be determined by screening, and (3) specifically when sample size is less than or equal to 250 that HC3 should be used.

```
summary(model1)
```

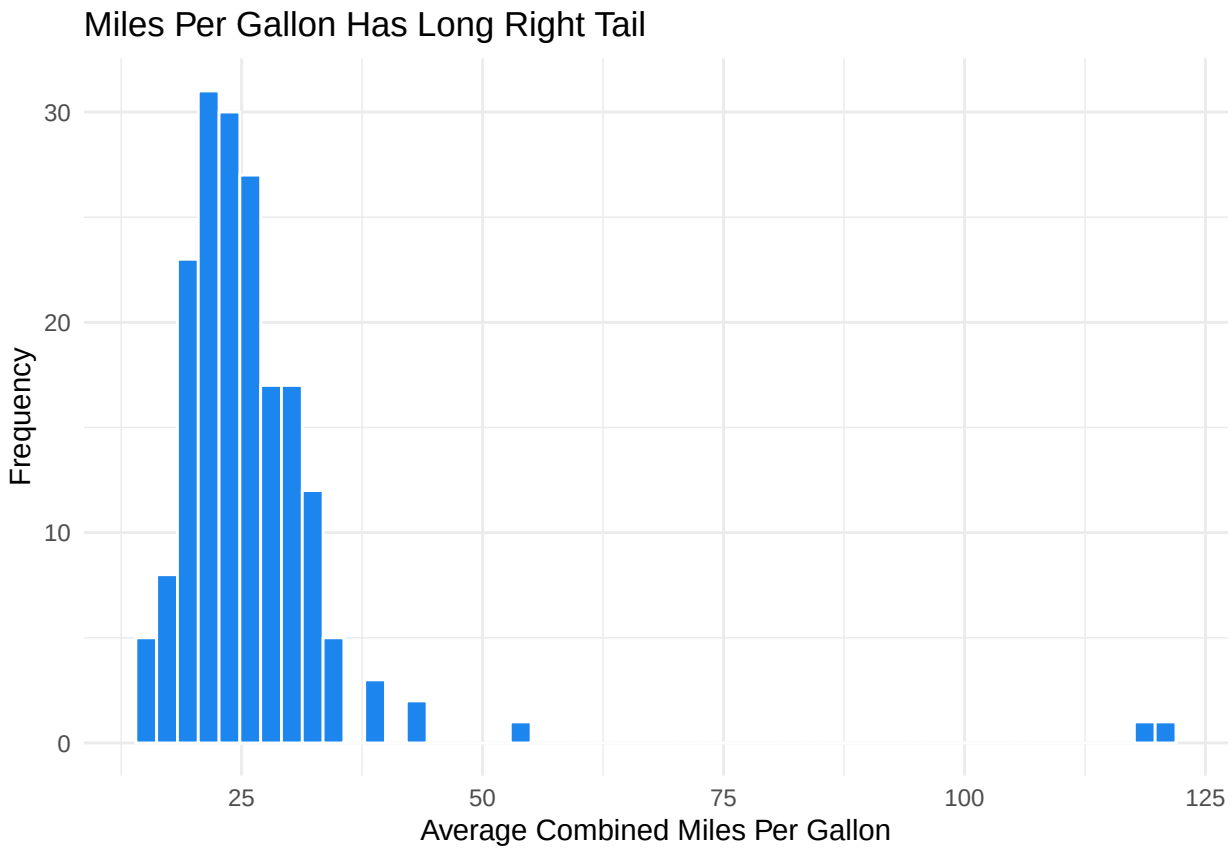
```
##
## Call:
## lm(formula = sales_unit_count_log ~ avg_hp, data = sale_by_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6116 -0.9926  0.0565  1.0550  3.8467
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.9710055  0.2866978  41.755  < 2e-16 ***
## avg_hp      -0.0058905  0.0009927  -5.934 1.47e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.393 on 181 degrees of freedom
## Multiple R-squared:  0.1629, Adjusted R-squared:  0.1582
## F-statistic: 35.21 on 1 and 181 DF,  p-value: 1.475e-08
```

The base linear model that only uses avg_hp(horsepower) to predict sales_unit_count(sales) is estimated as $log(\text{sales}) = 11.97 - 0.006 \cdot \text{horsepower} + \hat{\epsilon}$. Model 1 states that every increase in one unit of horse power decreases the sales by about 0.6%. Else, the R-squared and adjusted R-squared is only 0.1629 and 0.1582, which indicates that only using average horsepower as the explanatory variable is far from sufficient. We now introduce covariates to examine how robust average horsepower is to the inclusion of other variables.

Add Average Combined MPG

Fuel efficiency may influence the effect of horsepower so we introduce average combined miles per gallon as another covariate to create Model 2. Model 2 does not contain average city miles per gallon and highway miles per gallon since these variables introduce serious conlinearity without much expected benefit. The distribution, specification, and statistical summary of Model 2 follows:



We expect an inverse relationship between horsepower and fuel efficiency so let us specify the model and estimate its coefficients to examine the effect of miles per gallon.

$$\text{MODEL 2} \rightarrow log(\text{sales}) = \hat{\beta}_0 + \hat{\beta}_1 \text{horsepower} + \hat{\beta}_2 \text{mpg} + \hat{\epsilon}$$

```
summary(model2)
```

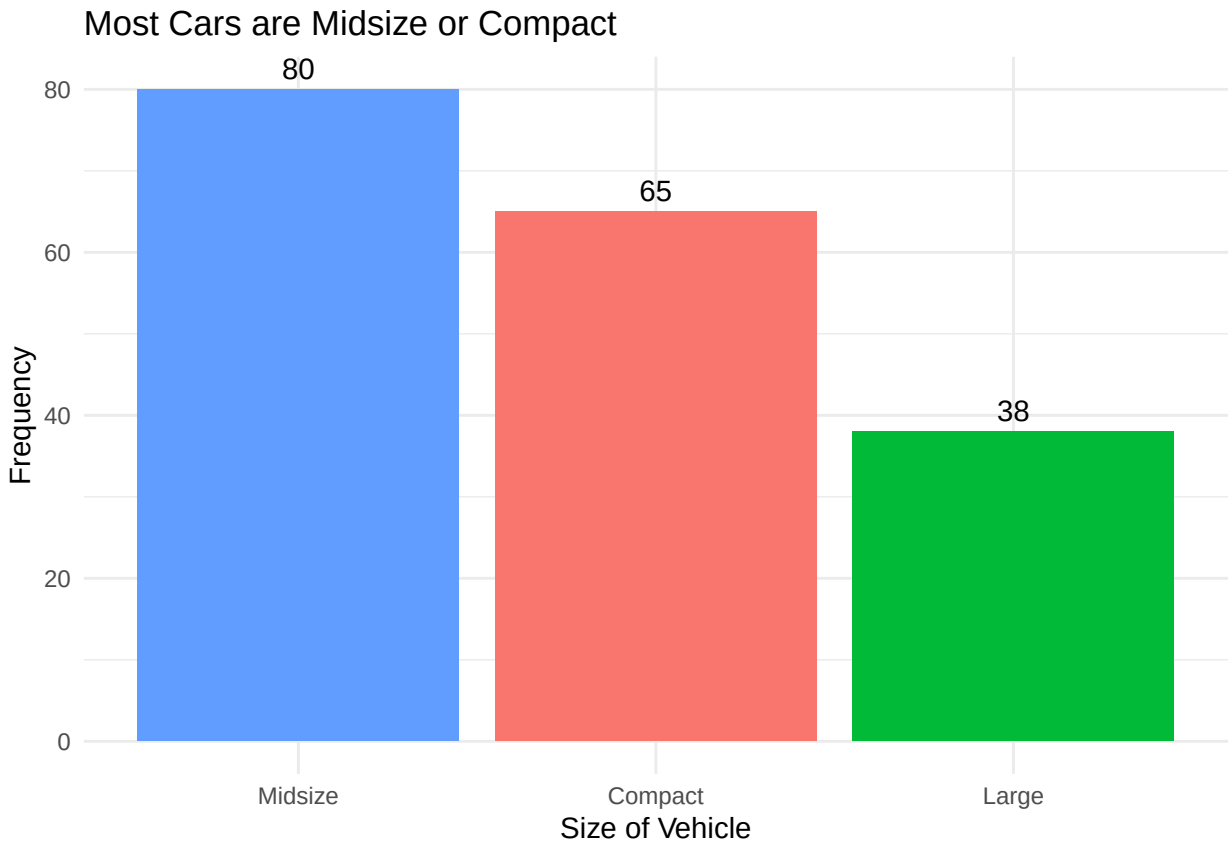
```
##
## Call:
## lm(formula = sales_unit_count_log ~ avg_hp + avg_cmb_mpg, data = sale_by_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6821 -0.9340  0.0737  1.0376  3.7818
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.506194   0.475430  26.305  < 2e-16 ***
## avg_hp      -0.006520   0.001086  -6.003 1.04e-08 ***
## avg_cmb_mpg -0.013995   0.009933  -1.409    0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.389 on 180 degrees of freedom
## Multiple R-squared:  0.172,  Adjusted R-squared:  0.1628
## F-statistic: 18.69 on 2 and 180 DF,  p-value: 4.202e-08
```

After variable avg_cmb_mpg(average combined MPG) has been added, the model changes to: $log(\text{sales}) = 12.506 - 0.007 \cdot \text{horsepower} - 0.014 \cdot \text{mpg} + \hat{\epsilon}$, where fuel efficiency is not statistically significant. When horsepower and average combined miles per gallon are included in the model, the R-squared and adjusted R-squared do not increase meaningfully in explaining the dependent variable.

Add Size

Looking to improve our model, we use size as a dummy variable. We have two dummy variables: midsize and large with compact as our baseline/omitted variable. The number of vehicles in the dataset for each size category is shown below:



You can see that most cars are compact or midsize. When bringing in the dummy variable of size, we create Model 3 and estimate coefficients.

$$\text{MODEL 3} \rightarrow log(\text{sales}) = \hat{\beta}_0 + \hat{\beta}_1 \text{horsepower} + \hat{\beta}_2 \text{mpg} + \hat{\beta}_3 \text{midsize} + \hat{\beta}_4 \text{large} + \hat{\epsilon}$$

```
summary(model3)
```

```
##
## Call:
## lm(formula = sales_unit_count_log ~ avg_hp + avg_cmb_mpg + size,
##     data = sale_by_model)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8857 -0.9410  0.0417  0.8996  3.1177
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.196118   0.461550  26.424  < 2e-16 ***
## avg_hp      -0.008671   0.001108  -7.824 4.36e-13 ***
## avg_cmb_mpg -0.005053   0.009499  -0.532 0.595386
## sizelarge    1.541484   0.307036   5.021 1.24e-06 ***
## sizemidsize  0.769242   0.227300   3.384 0.000878 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.305 on 178 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2616
## F-statistic: 17.12 on 4 and 178 DF,  p-value: 6.766e-12
```

The size of a car appears to impact sales. The estimated model is: $log(\text{sales}) = 12.200 - 0.009 \cdot \text{horsepower} - 0.005 \cdot \text{mpg} + 0.769 \cdot \text{midsize} + 1.541 \cdot \text{large} + \hat{\epsilon}$. All else equal, a car changing in size from compact to large will increase sales by a substantial amount. Given the nature of log-linear models, the exact increase is harder to measure compared to a linear-linear model. Our function `estimate_percent_impact()` reports that the estimated percent change in sales from an increase in size from compact to large in Model 3 is around 367%, albeit with a large confidence interval (156%, 753%). Midzsize cars also seem to drive sales more than compact cars, though to a lesser degree. Fortunately, the R-squared and adjusted R-squared in this model have increased to 0.2778 and 0.2616, so we can suggest that Model 3 performs better than our previous models.

The practical effect size measurements noted in the results section were calculated using a function inspired by University of Virginia Library[5]. Please see below for the implementation of percent change effect sizes for our log-linear models:

```
estimate_percent_impact <- function(coef, se) {
  # Takes coefficient and standard error of dummy variable in a log-linear
  # regression model and returns the estimated percent change in logged response
  # variable resulting from 1 unit change in non-logged explanatory variable as well as upper and
  # lower bound for 95% confidence interval
  estimate <- 100 * (exp(coef) - 1)
  lower_bound <- 100 * (exp(coef - (1.96 * se)) - 1)
  upper_bound <- 100 * (exp(coef + (1.96 * se)) - 1)

  return(c(estimate, lower_bound, upper_bound))
}

estimate_percent_impact(1.541484, 0.307036)
```

```
## [1] 367.1518 155.9197 752.7315
```

Add Luxury

Another dummy variable that we have available is the status of each vehicle as a luxury car, allowing us to write Model 4:

$$\text{MODEL 4} \rightarrow log(\text{sales}) = \hat{\beta}_0 + \hat{\beta}_1 \text{horsepower} + \hat{\beta}_2 \text{mpg} + \hat{\beta}_3 \text{midsize} + \hat{\beta}_4 \text{large} + \hat{\beta}_5 \text{luxury} + \hat{\epsilon}$$

Now using compact, non-luxury cars are our omitted variable we calculate our coefficients.

```
summary(model4)
```

```
##
## Call:
## lm(formula = sales_unit_count_log ~ avg_hp + avg_cmb_mpg + size +
##     IS_LUXURY, data = sale_by_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0170 -0.9468  0.1190  0.8377  2.9254
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.954152   0.462801  25.830  < 2e-16 ***
## avg_hp      -0.007443   0.001183  -6.292 2.39e-09 ***
## avg_cmb_mpg -0.001635   0.009427  -0.173 0.862523
## sizelarge    1.456008   0.303592   4.796 3.42e-06 ***
## sizemidsize  0.806341   0.223926   3.601 0.000412 ***
## IS_LUXURYY  -0.610645   0.228949  -2.667 0.008359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.283 on 177 degrees of freedom
## Multiple R-squared:  0.3057, Adjusted R-squared:  0.2861
## F-statistic: 15.59 on 5 and 177 DF,  p-value: 1.063e-12
```

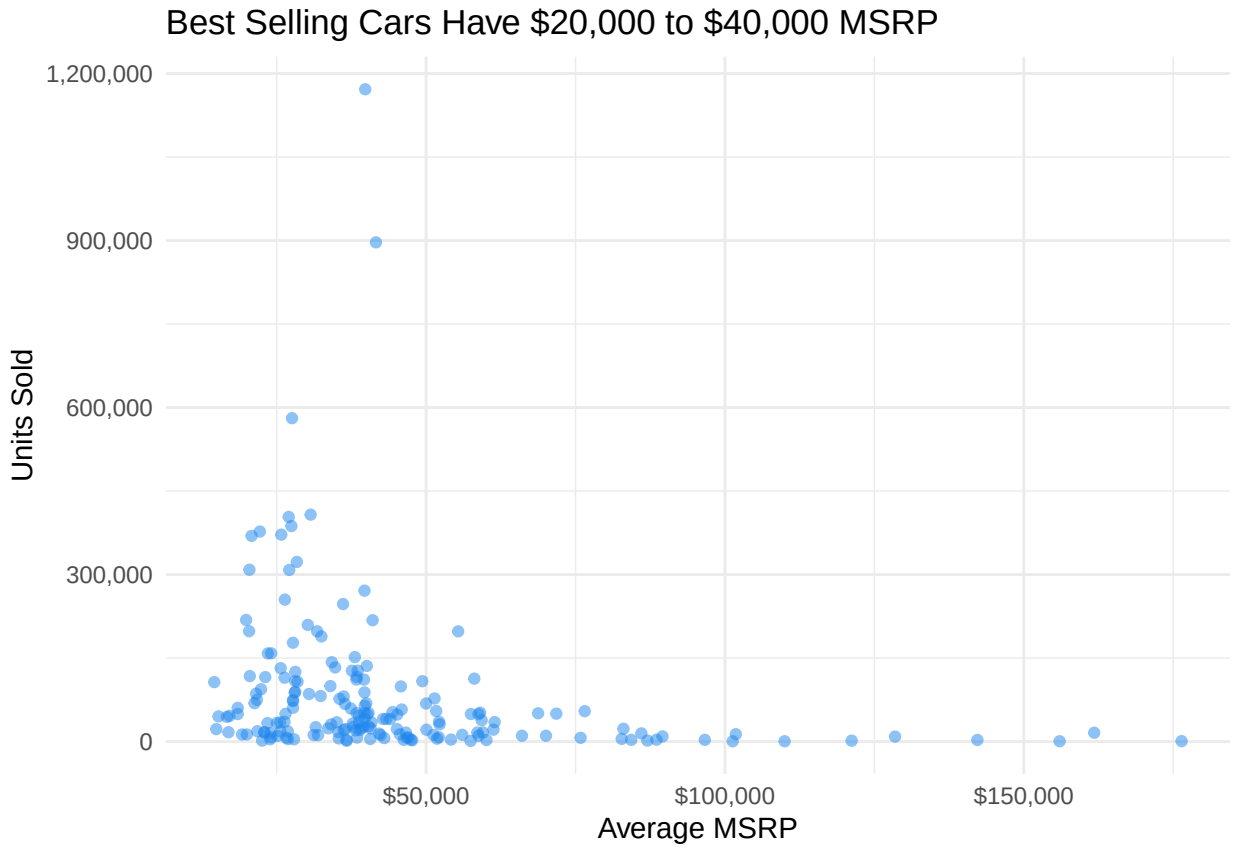
From Model 4, we can infer that the luxury brand also affect customers' willing toward buying cars. The coefficient for the dummy variable luxury is -0.61 and statistically significant. From the output of the model, we noticed that, compared those cars who are not luxury brand, the sales of luxury cars changes by -46% (-65%, -15%). Other statistically significant variables stay relatively stable. Happily, the R-squared and adjusted R-squared have both increased, which means that this model can better explain the dependent variable.

```
estimate_percent_impact(-0.610645, 0.228949)
```

```
## [1] -45.69995 -65.33311 -14.94780
```

Add MSRP

Up until this point, we have considered variables in our dataset that may affect sales without also considering each make/model's average MSRP, which we expect to have a substantive impact on sales. The best selling cars are within the range of around $20,000 to $40,000, as evidenced by the following scatterplot:



By incorporating average MSRP, we specify Model 5 and calculate its coefficients:

$$\text{MODEL 5} \rightarrow log(\text{sales}) = \hat{\beta}_0 + \hat{\beta}_1 \text{horsepower} + \hat{\beta}_2 \text{mpg} + \hat{\beta}_3 \text{midsize} + \hat{\beta}_4 \text{large} + \hat{\beta}_5 \text{luxury} + \hat{\beta}_6 \text{MSRP} + \hat{\epsilon}$$

```
summary(model5)
```

```
##
## Call:
## lm(formula = sales_unit_count_log ~ avg_hp + avg_cmb_mpg + size +
##     IS_LUXURY + avg_msrp, data = sale_by_model)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7102 -0.9352  0.0548  0.8695  2.8573
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.155e+01  4.801e-01  24.054  < 2e-16 ***
## avg_hp      -2.396e-03  2.230e-03  -1.074 0.284084
## avg_cmb_mpg  1.245e-03  9.334e-03   0.133 0.894046
## sizelarge    1.164e+00  3.182e-01   3.659 0.000334 ***
## sizemidsize  6.218e-01  2.309e-01   2.693 0.007774 **
## IS_LUXURYY  -3.517e-01  2.454e-01  -1.433 0.153604
## avg_msrp    -2.186e-05  8.236e-06  -2.654 0.008691 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.261 on 176 degrees of freedom
## Multiple R-squared:  0.3324, Adjusted R-squared:  0.3097
## F-statistic: 14.61 on 6 and 176 DF,  p-value: 1.666e-13
```

The resulting calculations give us a final model that helps us to understand the relationships in our dataset that explain sales:

$$\text{MODEL 5} \rightarrow log(\text{sales}) = 11.55 - 0.002 \cdot \text{horsepower} + 0.001 \cdot \text{mpg} + 0.62 \cdot \text{midsize} + 1.16 \cdot \text{large} - 0.35 \cdot \text{luxury} - 0.00002 \cdot \text{MSRP} + \hat{\epsilon}$$

In Model 5, horsepower, miles per gallon, and luxury are not statistically significant. Notably, our variable of interest, average horsepower, becomes statistically insignificant in Model 5 whereas in Model 4 is was statistically significant. Model 5 also reveals that evvery $1 increase in average MSRP would decrease sales by about 0.00219%. One other notable change is that when average MSRP is included in the model, the R-squared and adjusted R-squared have increased to 0.3324 and 0.3097, which is marked improvement over Model 1's Adjusted R-squared of 0.16. A summary of table of Models 1 - 5 is shown on the next page and an alternative path to Model 5, introducing MSRP first, is shown the following page to provide context on the interplay between MSRP and the other variables.

Table of Models

Table 3: Models 1 through 5

	Dependent variable:				
	sales_unit_count_log				
	(1)	(2)	(3)	(4)	(5)
avg_hp	−0.006***	−0.007***	−0.009***	−0.007***	−0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
avg_cmb_mpg		−0.014	−0.005	−0.002	0.001
		(0.010)	(0.009)	(0.009)	(0.009)
sizelarge			1.541***	1.456***	1.164***
			(0.307)	(0.304)	(0.318)
sizemidsize			0.769***	0.806***	0.622***
			(0.227)	(0.224)	(0.231)
IS_LUXURYY				−0.611***	−0.352
				(0.229)	(0.245)
avg_msrp					−0.00002***
					(0.00001)
Constant	11.971***	12.506***	12.196***	11.954***	11.548***
	(0.287)	(0.475)	(0.462)	(0.463)	(0.480)
Observations	183	183	183	183	183
R^2	0.163	0.172	0.278	0.306	0.332
Adjusted R^2	0.158	0.163	0.262	0.286	0.310
Residual Std. Error	1.393 (df = 181)	1.389 (df = 180)	1.305 (df = 178)	1.283 (df = 177)	1.261 (df = 176)
F Statistic	35.210*** (df = 1; 181)	18.694*** (df = 2; 180)	17.118*** (df = 4; 178)	15.588*** (df = 5; 177)	14.607*** (df = 6; 176)
Note:					*p<0.1; **p<0.05; ***p<0.01

To provide more context, we now introduce MSRP as the first covariate. Note how the inclusion of MSRP at the beginning of the model building process prevents horsepower, `avg_hp`, from becoming statistically significant while large (`sizelarge`), midsize (`sizemidsize`) dummy variables remain statistically significant when included.

Table 4:

	Dependent variable:			
	sales_unit_count_log			
	(1)	(2)	(3)	(4)
avg_hp	0.002	0.002	−0.002	−0.002
	(0.002)	(0.002)	(0.002)	(0.002)
avg_msrp	−0.00004***	−0.00004***	−0.00003***	−0.00002***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
avg_cmb_mpg		−0.004	0.0002	0.001
		(0.010)	(0.009)	(0.009)
sizelarge			1.143***	1.164***
			(0.319)	(0.318)
sizemidsize			0.564**	0.622***
			(0.228)	(0.231)
IS_LUXURYY				−0.352
				(0.245)
Constant	11.396***	11.541***	11.578***	11.548***
	(0.289)	(0.486)	(0.481)	(0.480)
Observations	183	183	183	183
R^2	0.274	0.274	0.325	0.332
Adjusted R^2	0.266	0.262	0.306	0.310
Residual Std. Error	1.301 (df = 180)	1.304 (df = 179)	1.265 (df = 177)	1.261 (df = 176)
F Statistic	33.924*** (df = 2; 180)	22.554*** (df = 3; 179)	17.016*** (df = 5; 177)	14.607*** (df = 6; 176)
Note:				*p<0.1; **p<0.05; ***p<0.01

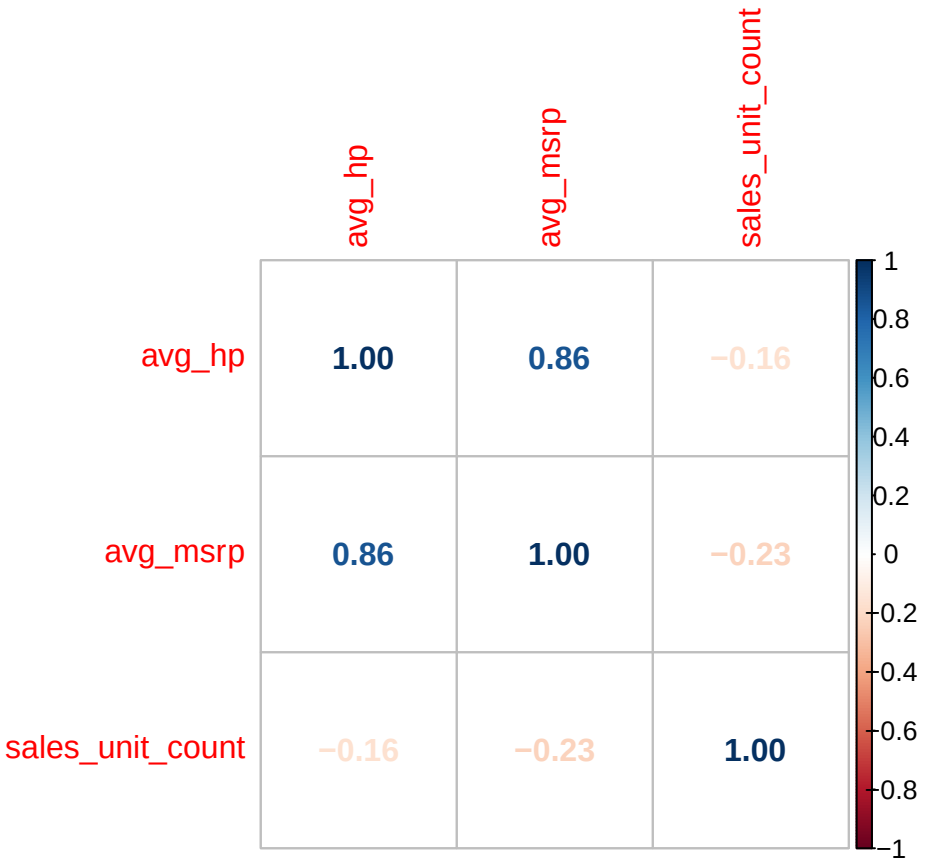
What Didn't Work

Models and/or specifications that we elected not to describe in detail in our the lab:

- Models in the linear-linear family like `lm(sales_unit_count ~ avg_hp + avg_cmb_mpg + size)`: Models that examined nominal sales as a function of nominal variables showed horsepower to be not statistically significant as soon as any covariates were introduced. We attempted several models of different specifications with different combinations of variables. These models had a maximum Adjusted R2 of 0.09, compared to the Adjusted R2 of 0.30 for non-msrp log-linear models. This fact may speak to the limited impact that horsepower seems to make on sales.
- Models in the linear-log family: no explanatory variable, such `avg_cmb_mpg` met the criteria for a valid log transformation so these types of models were not included.
- Models with both `avg_city_mpg` and `avg_highway_mpg`: These variables were highly colinear. We also wanted a proxy for fuel efficiency and we felt that `avg_cmb_mpg` alone was a good enough indicator for this concept.
- Models with interaction terms like `lm(sales_unit_count_log ~ avg_hp + size*avg_hp)` or `lm(sales_unit_count_log ~ avg_hp + is_luxury*avg_hp)`: No interaction term, such as `avg_hp:sizelarge`, was statistically significant. One might posit that once a person chose a particular size of vehicle that the amount of horsepower might then affects sales. For example, once a potential buyer decided they wanted a large vehicle, such as a truck or SUV, they might then want a truck or SUV with more horsepower. However, the lack of statistically significant interaction terms fails to provide evidence for this sort of preference.
- Models with transmission type as an explanatory variable: A failing of the data is that fact that 46 of the 184 models are coded as `various` for the transmission type, rather intractable category. These vehicles accounted for more than 29% of total sales, prohibiting us from reasoning clearly about transmission type's relationship to sales.
- Models that examined only cars with less than 210 horsepower: In the plot *First Glance: No Strong Relationship Between Horsepower and Sales*, one can see a vertical line of vehicles that have just more than 200 horsepower, indicating that manufacturers may be attempting to ensure their cars have at least 200 horsepower. For these lower powered cars, one could imagine that increased horsepower increased sales. Testing a model including only vehicles with horsepower less than 200, did not yield statistically significant results expected, and a discussion of such models is omitted.

4. Results

- **No strong evidence that horsepower increases sales**: Horsepower's effect on sales is stable and negative until MSRP is introduced, after which, horsepower is not statistically significant. We hypothesize the reason for this shift is that MSRP absorbs the effect of horsepower. MSRP and horsepower have a high correlation coefficient of 0.86. The relationship between horsepower and sales is less strong, as shown in the correlation heatmap below. In Models 1 - 4, as horsepower increases, sales decrease because horsepower and MSRP move together and higher prices decrease sales. Then, in Model 5, as MSRP is introduced, horsepower is no longer statistically significant. Examining Table 2 supports this view: In Table 2, average MSRP is the first covariate introduced, rendering horsepower statistically insignificant.



- **Size matters**: Our base case/omitted variable category is a non-luxury, compact car. There are huge sales boosts for companies that can produce cars larger than compact (trucks, SUVs, etc.) If not considering price, then larger cars sell more. These benefits are robust even to the introduction of MSRP. If considering our proxy for price, then large cars still sell more. The size increase from a a non-luxury compact to a non-luxury large car in Table 3, Model 5 is estimated at 220% increase, with a lower bound of 71% and upper bound of 487% for that estimate. We even found that interaction terms between horsepower and size were not statistically significant. As noted in the *What Didn't Work* section, there does not seem to be marginal effect of additional horsepower for midsize or large vehicles. Our models suggest that people simply seem to want larger vehicles, regardless of other variables.
- **Effect of luxury likely absorbed by MSRP**: In the same way that horsepower is related to MSRP, we imagine that the difference between Table 1 Model 4 when `IS_LUXURY` is statistically significant and Table 3 Model 5 when `IS_LUXURY` is not longer statistically significant is due to the colinearity between a car's status as "luxury" and its MSRP. In short, MSRP absorbs any potential effect of this dummy variable since luxury cars tend to be higher priced and sell at a lower volume.
- **Fuel efficiency does not affect sales**: No model that we tested, in Table 3 or Table 4 or in the "What Didn't Work" section, showed that fuel efficiency, as measured by average combined miles per gallon, has a statistically significant effect on sales volume.

Effect size calculations in percent (estimate, lower bound, upper bound):

Model 1 Percent change in sales caused by 1 unit increase in horsepower: -0.5982036, -0.7928403, -0.403185.

Model 5 Percent change in sales of non-luxury large vehicle compared to non-luxury compact: 220.2718563, 71.7242832, 497.3183292.

Model 5: Percent change in sales as result of $1 increase in MSRP: -0.002, -0.0039599, $-3.9999992 \times 10^{-5}$.

5. Limitations of Our Model

5a. Statistical Limitations

Large sample model assumptions:

The IID assumption is not met with confidence. Vehicle manufacturers mirror one anthers processes, product offerings, and marketing campaigns. Additionally, our treatment of the data may have further compromised the IID assumption. All large scale

sample models built on non-IID data are subject to suspicion since the underlying calculations the support their entire large scale sample model framework no longer operate as they need to. At this time we have no way to remediate this potential problem and readers should bear this problem in mind throughout our analysis.

Other statistical limitations/considerations:

- **Response vs Explanatory Variable Granularity**: Our publicly available sales volume data has sales listed by make and model but not by trim line. Our dataset containing features of each vehicles is listed by make and model and trim. Consequently, we are prevented from analyzing sales per make/model and trim, which might give us a better picture of the factors that affect sales. This issue also means that some variables in our dataset are difficult to manage. For example, data we have for transmission type is intractable. 29% of the vehicles in the dataset are listed as "trans_various," since different trim lines have various transmission options but we lack the capacity to consider the number of vehicles sold for each transmission type. We could not find publicly available data on sales by time line, so individuals with access to proprietary sales data will likely be able to address this limitation.
- **Size variable could be more specific**: Our dataset clumps vehicles into three groups: compact, midsize, and large. One could easily argue that the grouping used in this assignment collapses car make/models into too few categories. An analysis that groups vehicles into more specific classe such as SUVs, trucks, vans, subcompacts, etc. might illuminate the factors that influence sales better than our current models.
- **Manual assignment of 7 vehicle's sizes**: The dataset did not have clear size listed for 7 automobiles, so we hand coded the size for these 7 vehicles: Cadillac ATS (compact), Chevrolet Cruze (compact), Ford Transit (large), Honda Civic (compact), Infiniti q50 (midsize), Mazda 3 (compact), Subaru Impreza (compact). We do not anticipate that this attempt to "fill out" had a meaningful impact on the overall recommendations, however, we acknowledge it is an imperfect approach.

5b. Structural Limitations

There are some omitted variables. Engines of vehicles do not change nearly as quickly as attributes like MSRP so horsepower is relatively independent of other variables. Consequently, we are more concerned about how omitted variables could affect MSRP, which may be quite responsive to other variables.

Omitted Variable Bias

- *Brand Loyalty*: We will define brand loyalty as a customer's propensity to buy a new car of the same make as a previous car that they have purchased. Brand loyalty can play a large role in pricing. According to JD Power & Associates, 60% of Toyota and Subaru drivers that bought a Toyota or Subaru buy another Toyota or Subaru[6]. Our causal theory assumes that brand loyalty has a positive relationship with sales; increased loyalty causes an increase in sales. We also assume brand loyalty positively impacts MSRP; stronger brand loyalty allows a manufacturer to increase their MSRP. These two assumptions mean omitted variable bias for brand loyalty is positive. Since our coefficient for MSRP, however, is negative, the direction of the bias is towards zero. While we are concerned that the this omitted variable bias exists, we note that it is going towards zero, causing less concern. To resolve this omitted variable bias, we could attempt to join brand loyalty survey or manufacturer's data to our current dataset.
- *Style*: Some customers care about the interior/exterior appearance of their vehicle. We do not know the direction of this bias because the relationship between style, horsepower, and sales or style, MSRP, and horsepower is hard to assume. An approach to mitigating the omission of this variable would be to scrape car review articles from car websites and blogs, then run a sentiment analysis from the resulting corpus, analyzing appearance and style. We acknowledge that this strategy would be difficult, time-consuming, and potentially not yield meaningful information.
- *Demand for cars*: Overall demand for cars as the result of good economic health is another potential ommitted variable. "Economic health" could be operationalized to any number of variables: stock market performance relative to 10 year baseline average performance, employment rates, consumer spending, etc. For the moment, we will take consumer spending as a measure of economic health. If consumer spending increases then it could have an positive impact on MSRP. Spending could also positively increase sales, creating a positive bias. The direction of this bias would be toward from zero since the coefficient for MSRP with respect to sales is negative. This sort of bias could be addressed by including some measure of economic health in the model, such as the aforementioned measure of consumer spending.

The issues listed above are not meant to be understood an an exhaustive list of omitted variables. Other factors such as the actual price that consumers pay (not MSRP), marketing campaigns, etc. will also affect are model but are not captured within it.

Reverse Causality

Sales and MSRP: There likely exists reverse causality between sales and MSRP. The exact nature of this relationship is hard to determine. Increased prices in general lead to lower sales. However, increased sales may lead to higher prices if manufacturers are confident they can earn profits by maintaining sales volume on higher prices. In other cases, an increase in sales might indicate

that a pricing strategy is working and prices might drop even further. We can be certain only that prices and sales are linked but are unable at this time to say exactly how.

Horsepower and MSRP: An increase in horsepower may put upward pressure on MSRP since powerful motors are more likely to cost more to produce. Manufacturers would then increase MSRP to cover the cost of added horsepower. At the same time, MSRP may influence horsepower. Many auto manufacturers create tiers of vehicles at different price points in order to target different segments of the market. If they first choose a price then choose an engine that allows them to profit from cars in that price range, we can see that horsepower affects price while price also affects horsepower. Statistically, the result of this issue of reverse causality is that the coefficients and standard errors of horsepower and MSRP might not be trustworthy. Structurally, this means a core assumption of the structural equation modeling process is broken so the model itself becomes less persuasive. Our data, on the other hand, suggests that price is such a strong factor affecting sales that the effect of this particular case of reverse causality may be limited.

Omitted variable bias and reverse causality are cause for concern and potentially render our models inoperable. At this time, however, we do not have the means to address these concerns and leave it to the reader to gauge the severity of these violations of the structural equation models.

Conclusion

Recommendations

- **Take no action to increase horsepower**: When omitting MSRP, increases in horsepower have an negative impact on sales. As soon as one includes MSRP, horsepower fails to be statistically significant. This effect may be a result of horsepower being absorbed by a statistically significant, practically significant, and highly colinear variable, MSRP. Furthermore, that our models managed only an Adjusted R^2 of around 0.33 suggests that important variables affect sales that are not captured in our models. The introduction of currently omitted variables might decrease the already questionable effect of horsepower. With these considerations in mind, we do not recommend auto manufacturers to invest in ways to increase the horsepower of their vehicles.
- **Bigger is better**: The effect of the dummy variable for midsize and large vehicles is strong and statistically significant even when considering MSRP. We have some evidence to show that we can reject the hypothesis that vehicle size does not affect sales. We recommend that manufacturers who seek to increase sales volume in the United States consider focusing on large, and to a lesser extent, midsize vehicles. Our data supports the decision that Ford made to stop producing and selling sedans in the United States in 2020[7]. It may be the case that bigger is in fact better.
- **Lower prices**: While obvious, the fact that higher prices tends be associated with a decrease in sales is still important. Finding ways to reduce prices will likely increase sales.

Further Exploration

- **Get more or better data**: Get data to account for other variables that might explain sales, such as product loyalty, economic activity, government subsidies/taxes, advertisement spending, etc to improve performance of model. Likewise, obtaining more granular data on transmission type or specific sales/feature information for each make/model trim line would help.
- **Link sales volume to profit**: Luxury cars earn large marginal profit on fewer sales while non-luxury cars achieve profitability through lower margins and higher volume. Another next step might be to link sales to profit to better inform business decisions.

[1] https://www.statista.com/statistics/566098/research-development-spending-automotive-industry-worldwide/ [2] https://www.autoevolution.com/news/mitsubishi-discontinues-i-miev-in-the-united-states-no-replacement-planned-119591.html [3] Trying to be punny. [4] https://jslsoc.sitehost.iu.edu/files_research/testing_tests/hccm/99TAS.pdf (Shared by Alex in 203 Slack channel) [5] https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/ [6] https://www.businessinsider.com/car-buying-brands-most-loyal-customers-automotive-sales-loyalty-subaru-2020-7?op=1#2-toyota-19 [7] https://www.cnet.com/roadshow/news/ford-fusion-production-discontinued-sedan/