Prices for Hybrid Vehicles: What Features Affect Prices?

Shuangjie Zhang[1], Yuqing Zhang[1]
UCSC Statistical Science Department[1]
https://github.com/Zsj950708/STAT-204-Final-Project

Abstract

Multiple linear regression model is widely used to find the relationship between the response variable and explanatory variables. This paper uses a multiple linear regression model to describe the relationship between the suggested retail price and the features of hybrid vehicles. After balancing the trade-off between accuracy and model complexity, we come up with a final model consisting of acceleration rate, mile-per-gallon, three representative car brands, and one representative car class. For model diagnostics, we check the normality assumption and identify one outlier with the mean-shift outlier model and one high leverage point. In summary, we conclude that a set of hybrid car features will affect the suggested retail price.

KEY WORDS: Multiple Linear Regression, AIC, BIC, Lasso, Mean-shift Outlier Model

1. Introduction

Consider buying a car? We may all have the painful moments struggling with tons of search results when we only have a vague idea about what kind of car we want. The only thing for sure is probably the budget. In this project, we consider hybrid electric cars and provide a solution by answering the question: what features of a hybrid vehicle affect its price. With such information, buyers can balance between the features that they care most about and the budget. For example, we find that acceleration rate is strongly positively related to the price of a hybrid electric vehicle, while model year does not affect the price as much as we commonly believed. A buyer may then decide that he wants the newest model with lower acceleration rate, so that he gets the most stylish look and well controls the budget at the same time.

We focus the study on hybrid electric vehicles (HEVs). Since the first release of Toyota Prius in 1997, HEVs "have been brought to the automobile market around the world" in an effort to "address environmental and fuel-dependency concerns" [1]. We will promote this effort by exploring the explanatory variables for prices of HEVs and facilitating purchase decisions for HEVs.

1.1 Hybrid Electric Vehicles Data Set

Our study is based on a data set provided in "*Comparing technological advancement of hybrid electric vehicles (HEV) in different market segments*" [1]. The paper incorporated EPA (United States Environmental Protection Agency) database, technical report, product manual, and other sources to form a data set on technical attributes of HEVs. A total of 9 variables and labels are selected.

1.2 Variables of Interest

Variables of interest from the HEV data set are: MSRP, Acceleration rate, MPG, Max of MPG and MPG equivalent, Vehicle class, and Model year.

MSRP: Manufacturer's suggested retail price, the response variable of our study. Since vehicles in the data set were "from different countries and released in different years," Lim et al. took three steps to convert the actual MSRP into 2013 U.S. dollar value [1].

1. The vehicle's MSRP in the year of release was found through manufacturers' website or car review websites;
2. If the MSRP was in currency other than U.S. dollars, the value was converted to U.S. dollars using the exchange rate at the year of release provided by OANDA Corporation;
3. If the vehicle was released before 2013, the MSRP was inflated to present value by multiplying the ratio of Consumer Price Index (CPI) of 2013 to the historical CPI. The CPI values were obtained from the Bureau of Labor Statistics.

Acceleration rate: the average acceleration rate (kilometers/hour/second) of a vehicle as it goes from 0 to 100 km/h.

Max of MPG and MPG equivalent: MPG, or mile-per-gallon, is a measure of fuel economy for conventional HEVs. MPGe, or mile-per-gallon equivalent, is developed by EPA to estimate the fuel economy for plug-in HEVs. MPGe is calculated by approximating 1 gallon of gasoline to 33.7 kilowatt/hour of electric energy. The variable mpgmpge takes the maximum of the two, so both conventional and plug-in HEVs can be taken into account in modeling.

Vehicle class: Lim et al. adopted EPA's criteria and grouped HEVs into 7 classes: two-seaters (TS), compact (C), midsize (M), large (L), sport utility vehicle (SUV), minivan (MV), and pickup truck (PT) [1].

In addition to the variables provided by the HEV data set, we are also interested in the potential effects of vehicle brand, company, and origin country of the brand. So we map each car model to create 3 new variables. In summary, variables of interest are shown in Table 1.

Table 1: Data Variables Insight

Variables	Description	Data Type	Source
msrp	Manufacturer's suggested retail price in 2013	number	HEV data set
vehicle	Name of car models	factor	HEV data set
year	Model release year	integer	HEV data set
accelrate	Acceleration rate in km/hour/second	number	HEV data set
mpg	Fuel economy	number	HEV data set
mpgmpge	Max of mpg and mpge	number	HEV data set
carclass	Model class	factor	HEV data set
brand	Car brand	factor	Car review websites
company	Car company	factor	Car review websites
country	Origin country	factor	Car review websites

1.3 Prior Analyses

The HEV data set was initially brought about to explore the relationship between technological advancement and HEV market segments [1]. Msrp was "considered as a reasonable proxy for manufacturing cost", thus taken as an input variable for a technological forecasting model. In comparison, we use msrp as the response variable and manage to build an explanatory model in this study.

Other prior works include analyses conducted by students of the University of Toronto and the University of Illinois at Urbana-Champaign for homework assignments [2] [3]. Only a few variables were used in isolated linear regression practices. In contrast, we use the entire data set, as well as additional variables that can be mapped from existing data, to find the best explanatory model for suggested retail price of hybrid cars.

2. EDA

The raw data comes in a csv file that we downloaded from a database of University of Florida. The raw data contains 153 observations for 9 variables. There is no missing value. We first map vehicle brand, company, and country using model names, then remove irrelevant variables including vehicle ID, model name, and car class ID.

2.1 Pairwise Correlations

We use ggpairs function to observe the distribution of each variable and to explore the pairwise relationships in the HEV data set. For this analysis, we temporarily exclude brand (29 levels), company (21 levels), and country due to a large amount of levels in these factors. In Figure 1, we plot the pairwise correlations.

From Figure 1, we find that:

For individual variables:

- Model year ranges from 1997 to 2013 with more data in recent years;
- Msrp is right-skewed with some extremely large values, suggesting log transformation;
- Acceleration rate is close to normal distribution;

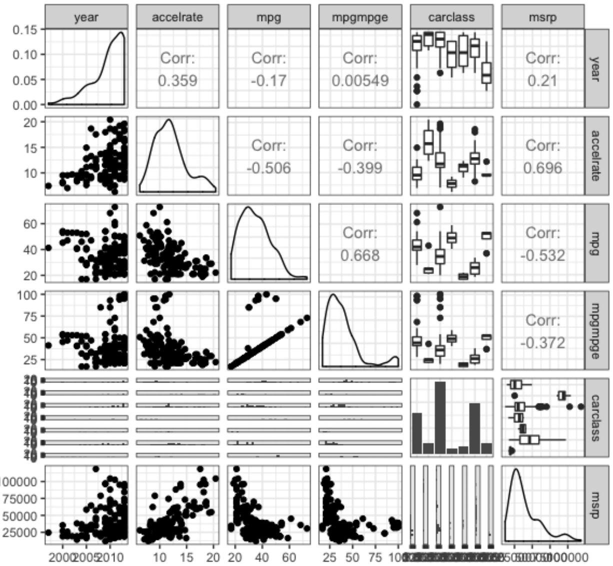

Figure 1: Pairwise Correlations

- Some factor levels only have a few observations, calling for attention in future analysis.

For pairwise relationships:

- The correlation coefficients show that acceleration rate and mpg may have a strong impact on msrp respectively;
- From the scatterplots, we see a potential linear relationship between acceleration rate and msrp, while mpg and mpgmpge need transformation to be model inputs;
- Mpg and mpgmpge are similar in distribution and have a strong correlation as expected.
- The distributions of msrp in some car classes and countries are very different from other groups. However, since we plan to work on a multiple regression model instead of ANOVA, car class and country may or may not be significant explanatory variables.

With these preliminary analyses, we decide to transform msrp, mpg and mpgmpge to log form, respectively.

Table 2: Full Model ANOVA Table

	Response: log(msrp)
year	48.097^{***}
accelrate	288.772^{***}
log(mpg)	50.914^{***}
log(mpgmpge)	2.218
carclass	7.980^{***}
brand	5.682^{***}

Note: $^{*}p < 0.05;^{**} p < 0.01;^{***} p < 0.001$

2.2 Full Model

We first fit a full model with log(msrp) as the response variable and all the others as explanatory variables. Table 2 displays the F values of full model.

We notice that company and country are not defined in this model due to singularities in data. We see from the ANOVA table that all variables except for log(mpgmpge) contribute to log(msrp). Residuals are close to normal distribution with no particular pattern in spite of a few outliers (residual plots for the full model are not shown). However, a closer look into coefficients reveal many insignificant levels of car class and brand. Next, we use various methods to select variables and build a better model.

3. Method

3.1 Multiple Linear Regression

We propose to use multiple linear regression on log(msrp). Here we define the general form of our regression. $X_{n\times(p+1)}$ is the design matrix that includes all covariates and the first column with intercept 1. Although we make some transformation afterwards, the general form is still a multiple linear regression model. Y is our response variable with dimension $n \times 1$. βs are the parameters we need to estimate with dimension $(p+1) \times 1$. ϵ is the error term, which follows a normal distribution.

$$Y_n = X\beta + \epsilon, \epsilon \sim N_n(0, \sigma^2 I_n)$$

In order to use the multiple linear regression model, we need to check whether the data fits the normality assumption. In goodness of fit section, we check the normality.

3.2 AIC/BIC

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are two conventional methods to compare various models. AIC was first formulated by Hirotugu Akaike [4] and BIC was developed by Gideon E. Schwarz in his 1978 paper [5]. Both AIC and BIC take into account the trade-off between the goodness of fit of the model and the complexity of the model.

$$\begin{aligned}\text{AIC} &= -2\log\text{likelihood} + 2k \\ \text{BIC} &= -2\log\text{likelihood} + k \times \log(n)\end{aligned}$$

Here k represents the number of parameters included in the model. And n is the number of observations in the data. Usually, BIC allows more penalty on the number of selected parameters. So BIC tends to select fewer variables than AIC. Smaller AIC and BIC values indicate a larger likelihood and better model.

3.3 Lasso

Lasso was first introduced to the field of statistics in 1996 by Robert Tibshirani [6]. Lasso introduces the penalty term in the context of least squares. Since it uses L_1 norm, Lasso can shrink some coefficients to 0, which achieves the goal of variable selection.

$$\begin{aligned}&\min\ \frac{1}{n}||Y - X\beta||_2^2 \\ &s.t.\ ||\beta||_1 \le t\end{aligned}$$

Before performing Lasso, we also implement stepwise selection combined with AIC/BIC to select variables. Although Lasso can achieve all the goals that stepwise does, basic Lasso has its drawbacks. Basic Lasso treats each level of factor variables as individuals. As a result, the method either selects all levels into the model or excludes the factor altogether. To make up for this, we identify significant representative levels in the categorical variables in Phase 2 of model comparison.

4. Variable Selection

4.1 Stepwise AIC/BIC

We first run `stepwise` function combined with AIC and BIC to select variables automatically. AIC chooses year, acceleration rate, log(mpg), log(mpgmpge), carclass, and brand. BIC only sticks to the acceleration rate and log(mpg). This result coincides with what is suggested in the method section. Since 153 observations were collected in the data set, $\log(n) = \log(153) \approx 5$, BIC allows for more penalty on the number of coefficients.

Not surprisingly, acceleration rate and mpg are two important car features affecting car price. They are two key features people look at when they make purchase decisions. Surprisingly, the country variable is not significant. Although there are only five categories, much fewer than 26 categories of the brand, the result still shows no difference in mean car price among different countries.

Because there is a big difference between the results from AIC and BIC, we perform the second round of variable selection with Lasso.

4.2 Lasso

We first implement Lasso regression on the full model by `gamlr` function. The result is listed in the first column of Table 3. All the coefficients of country variables are 0 except Korea, which agrees with the result from stepwise analysis. Furthermore, coefficients of some brand like

Cadillac or Jeep are not shrunk to 0, but coefficients of other brands like Audi or BYD are shrunk to 0. For the carclass variable, only two levels are non-zero.

Since there are fewer levels in the country variable than in the brand variable, we drop the brand variable in the second run of Lasso to see whether the brand variable has masked the effect of other variables. The second Lasso (second column of Table 3) shows that all coefficients of the country variable are 0. This step further convinces us that the effect of the country variable is not significant, which agrees with the result from stepwise analysis.

In the last step, we drop the country variable and add back the brand variable. From the third column of Table 3, we conclude that year is always shrunk to 0. That is, in our model, the year variable does not affect the suggested retail price of hybrid cars. A possible explanation is that manufacturers name the price based on manufacturing cost, functionality of car, etc. Model year may be more important in marketing, so dealers may set a newer car model at a higher market price, but that is not the case for manufacturer's suggested price.

The max of mpg and mpge (mpgmpge) is expected to be not significant in the model. The correlation between mpg and mpgmpge is 0.834 after log transformation. In order to eliminate multicollinearity, the Lasso model excludes the mpgmpge variable.

For the car class variable, the third Lasso result is similar to the last two. Namely, coefficients of class Large, Midsize, Minivan are significant. Moreover, for the brand variable, some levels including Jeep, Cadillac, and Mercedes-Benz have non-zero coefficients.

In summary, the remaining variables after two rounds of selection are acceleration rate, log(mpg), carclass, and brand. In the next step, we compare a series of models with these four variables, and identify representative car class and brand to find the final model.

5. Model Comparison

5.1 Phase 1

In Phase 1, we consider each variable as a whole and build four models.

As shown in Table 4, we first fit LM1 as the full model following variable selection. However, only acceleration rate, two car class levels, and two brand levels have significant coefficients. In LM2, we keep only acceleration rate, log(mpg), and their interaction. LM2 turns out to be a good model in terms of coefficients, while the adjusted R-squared is only 0.571 and not satisfying.

So we add back carclass in LM3. This time we get two car class levels in addition to acceleration rate and the interaction term as significant variables. If we substitute carclass with brand, as in LM4, we get significant coefficients for acceleration rate, log(mpg), their interaction, and three brands.

Comparing AIC and BIC values for these four models, we find that AIC prefers LM4 while BIC prefers LM2.

Table 3: Three Lasso Results

	Variables	Lasso1	Lasso2	Lasso3
1	year	0.00	0.00	0.00
2	accelrate	0.06	0.07	0.07
3	mpg	-0.22	-0.34	-0.24
4	mpgmpge	0.00	0.00	0.00
5	carclassL	0.15	0.22	0.14
6	carclassM	-0.05	0.00	-0.04
7	carclassMV	0.00	0.00	0.03
8	carclassPT	0.00	0.00	0.00
9	carclassSUV	0.00	0.00	0.00
10	carclassTS	0.00	-0.06	0.00
11	brandAudi	0.00		0.00
12	brandBesturn	-0.01		0.00
13	brandBMW	0.16		0.07
14	brandBuick	0.00		0.00
15	brandBYD	0.00		0.00
16	brandCadillac	0.42		0.30
17	brandChevrolet	0.00		0.00
18	brandChrysler	0.00		0.00
19	brandDodge	0.00		0.00
20	brandFord	0.00		0.00
21	brandGMC	0.07		0.00
22	brandHonda	-0.23		-0.19
23	brandHyundai	0.00		0.00
24	brandInfiniti	0.00		0.00
25	brandJeep	-0.22		-0.05
26	brandKia	0.00		0.00
27	brandLexus	0.08		0.02
28	brandLincoln	0.00		0.00
29	brandMazda	0.00		0.00
30	brandMercedes-Benz	0.27		0.18
31	brandMercury	0.00		0.00
32	brandNissan	0.00		0.00
33	brandPeugeot	0.00		0.00
34	brandPorsche	0.12		0.03
35	brandSaturn	-0.10		-0.02
36	brandToyota	0.00		0.00
37	brandVauxhall	0.00		0.00
38	brandVolkswagen	0.00		0.00
39	country_combinedJapan	0.00	0.00	
40	country_combinedKorea	-0.01	0.00	
41	country_combinedOthers	0.00	0.00	
42	country_combinedUSA	0.00	0.00	

Table 4: Phase 1 Model Comparison

		Response: log(msrp)			
		LM1	LM2	LM3	LM4
accelrate		0.052^{***}	0.480^{***}	0.392^{**}	0.541^{***}
log(mpg)		-0.095	0.806^{*}	0.674	1.430^{***}
carclass	L	0.292^{*}	-	0.383^{**}	-
	MV	0.310^{*}	-	0.161	-
	TS	-0.006	-	0.283^{*}	-
brand	Cadillac	0.711^{*}	-	-	0.922^{**}
	Jeep	-0.851^{*}	-	-	-0.683^{*}
	Mercedes-Benz	0.451	-	-	0.560^{*}
accelrate:log(mpg)		-	-0.116^{***}	-0.092^{*}	-0.140^{***}
constant		10.136^{***}	6.703^{***}	7.184^{***}	4.779^{***}
AIC		-413.329	-346.899	-353.330	-427.684
BIC		-301.203	-334.777	-323.026	-330.710
Adj R-Squared		0.768	0.571	0.604	0.784
Note:			$^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$		

Notice that LM4 has more variables than LM2, the results are well expected. LM2 and LM4 become candidate models.

5.2 Phase 2

One common issue for the models in Phase 1 is that many levels of car class and brand are not significant. In Phase 2, we identify significant levels of car class and brand to improve the explanatory model. Specifically, we encode the three significant car classes (Large, Minivan, Two-Seater) and three significant brands (Cadillac, Jeep, Mercedes-Benz) into 6 dummy variables with "1" if "Yes" and "0" if "No".

In LM5, we add in the three brands and get all coefficients significant. Then we continue to add in the three car classes. Two-Seater is the only car class that has significant coefficient. So we remove the two insignificant car classes and finally get LM7. A comparison of LM2, LM4, and LM5 through LM7 are shown in Table 5.

Again, AIC, BIC, and adjusted R-Squared give us different results. AIC and adjusted R-Squared favor LM4, which uses acceleration rate, log(mpg), and all levels of brand to estimate log(msrp). BIC favors LM7, which incorporates acceleration rate, log(mpg), one car class, and three brands. Reflect that the goal for this study is to identify all relevant explanatory variables to best estimate suggested retail price for hybrid cars. Therefore, we decide that LM7 better serves this purpose. Firstly, all coefficients in LM7 are significant, implying that all the covariates in LM7 make meaningful contributions to explaining the variation of log(msrp). Secondly, AIC and adjusted R-Squared of LM7 are close to that of LM4, the "best" model they chose. Thirdly, we justify that BIC is a more reliable evaluation of model in our study. Remember that the penalty term in BIC for our data set is big ($\log(153) \approx 5$). Given such big penalty, LM7 still prevails other models in terms of BIC, suggesting a significant increase in likelihood. Therefore, LM7 is the winning model.

5.3 Check for Interaction

Next, we want to double check for any meaningful pairwise interaction between the covariates of LM7. We use $log(msrp) = (accelrate + log(mpg) + brandCadillac + brandJeep + brandBenz + classTS)^2$ as the full model and pass it into `step` function. In the end, only the interaction between acceleration rate and log(mpg) is significant, confirming LM7.

6. Goodness of fit

6.1 Normality Assumption

After fitting a model, it is essential to check whether the model satisfies the normality assumption. A QQ(quantile-quantile) plot is a probability plot, which compares two probability distributions by plotting their quantiles against each other. In the regression setting, we plot the empirical quantiles against the normal theoretical quantiles. If the empirical quantiles are close to the theoretical ones, it means that our model satisfies the normality assumption.

From Figure 2, the QQ plot shows that most data points are listed on the line. It convinces us that our model satisfies the normality assumption. There are still some points, like case 21 and case 36, that are off the line. In the outlier test (section 6.3), we identify that case 21 is indeed an outlier. For case 36, it is a 2008 Toyota Crown. Among all Toyota hybrid car models, Crown has the highest retail price, probably due to its high mpg. However, Crown also has the fourth-smallest acceleration rate. Since our model considers acceleration rate as positively related to price, the estimated price of Crown is obviously lower than recorded price. Nevertheless, as we will see in the outlier test section, case 36 is not identified as an outlier.

Table 5: Phase 2 Model Comparison

		Response: log(msrp)				
		LM2	LM4	LM5	LM6	LM7
accelrate		0.480***	0.541***	0.491***	0.457***	0.522***
log(mpg)		0.806*	1.430***	0.965**	0.909*	1.128**
carclass	L	-	-	-	0.201	-
	MV	-	-	-	0.147	-
	TS	-	-	-	−0.295**	−0.313**
brand	Cadillac	-	0.922**	0.885***	0.888***	0.898***
	Jeep	-	−0.683*	−0.694*	−0.696*	−0.700*
	Mercedes-Benz	-	0.560*	0.472***	0.390**	0.477***
accelrate:log(mpg)		−0.116***	−0.140***	−0.118***	−0.109***	−0.127***
constant		6.703***	4.779***	6.047***	6.286***	5.513***
AIC		−346.899	−427.684	−375.719	-382.568	−382.577
BIC		−334.777	−330.710	−354.506	−352.264	−358.333
Adj R-Squared		0.571	0.784	0.651	0.673	0.669
Note:				$^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$		

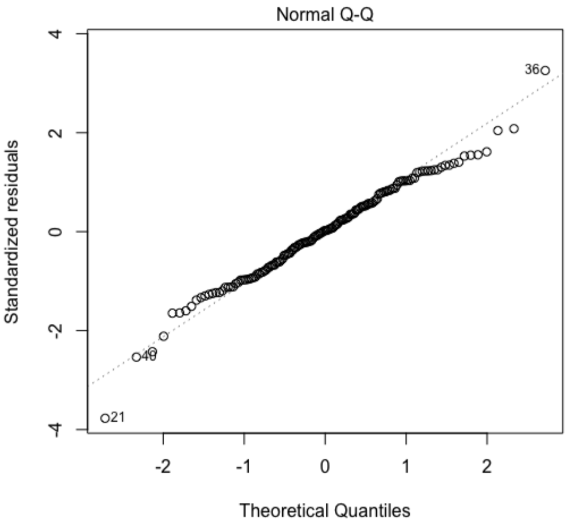

Figure 2: QQplot of final model

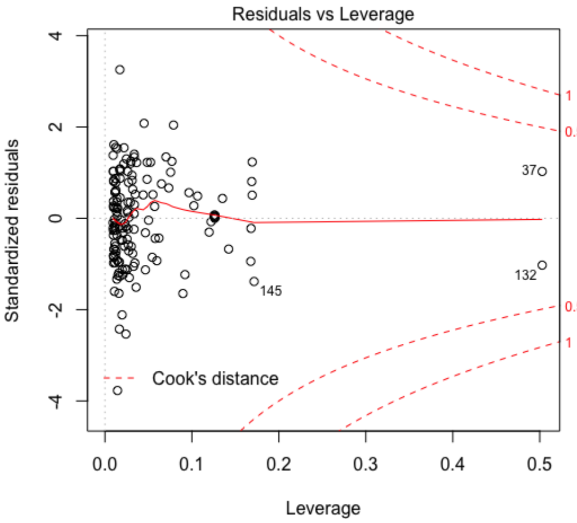

Figure 3: Leverage of final model

6.2 High Leverage Point

Hoaglin and Welsch (1978) suggest a direct use of the leverage score as a diagnostic to identify 'high-leverage points' [7]. Leverage score is typically used in the linear regression model, which measures how far away one observation is from all other observations. The motivation behind this suggestion is based on the representation:

$$\hat{y_i} = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$
$$H = X(X^TX)^{-1}X$$
$$h_{ii} = [H]_{ii}$$

Since hat matrix H is an orthogonal matrix, we can get that leverage is bounded between 0 and 1. If h_{ii} is large and close to 1, then the fitted value y_i is dominated by the i th observation y_i.

From the Cook's Distance plot and warning from R, we find that in our final model, there is only one observation with large leverage one. Referring to the original data set, we find that there is only one observation in that subgroup. That leads to high leverage. A lesson learned for future work is to consider combining levels for categorical variables.

6.3 Outlier test

The mean-shift outlier model is used to check whether i-th observation is suspected as being an outlier. The mean-shift outlier model is given by:

$$Y = X\beta + \phi d_i + \epsilon$$
$$\epsilon \sim N(0, \sigma^2 I_n)$$

where d_i is an n-vector with i-th element equals one, and all other elements equal zero. Nonzero values of ϕ implies the i-th observation is an outlier. Two hypotheses are brought up here:

$$H_0 : \phi = 0$$
$$H_1 : \phi \neq 0$$

It can be shown that the t-test for testing $H_0 : \phi = 0$ is equal to externally standardized residuals t_i. Externally standardized residuals t_i is given by:

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}/(1 - h_{ii})}}$$

where $e_{(i)}$ is predicted residual without using case i, $\hat{\sigma}^2_{(i)}$ is the estimated variance using data with case i deleted and h_{ii} is the leverage score for case i.

The mean-shift outlier model can be extended to include multiple d_i and d_j to test that case i and j are both outliers. When implementing multiple hypothesis testings, we use Bonferonni Correction to adjust the p-value. Bonferroni multiplicity adjustment compare each p-value to α/n and reject the null hypothesis (point is not an outlier) if the p-value is less than α/n. Bonferroni correction controls for familywise error rate (FWER), which is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned} &\text{FWER} = P\{\bigcup_{i=1}^{n_0}(p_i < \frac{\alpha}{n})\} \leq \sum_{i=1}^{n_0} P(p_i < \frac{\alpha}{n}) \\ &\text{FWER} \leq n_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} = \alpha \end{aligned}$$

One benefit of Bonferroni correction is not requiring any assumptions about dependence among the p-values or about how many of the null hypotheses are true.

In our final model, we identify one observation as an outlier with adjusted p-value smaller than 0.05 (Table 6). That is case 21, 2005 Honda Accord. Honda Accord has the highest acceleration rate among all Honda brand car models. From our final model, acceleration rate has a positive effect on the suggested retail price. So our model predicts higher price for Honda Accord than the recorded price.

Table 6: Outlier T-test result

	rstudent	unadjusted p-value	Bonferroni p
21	1-3.958069	0.00011826	0.017976

7. Conclusion

The final model we choose for estimating the manufacturer's suggested retail price of hybrid cars include acceleration rate, log(mpg), brandCadillac, brandJeep, brandBenz, classTS, and the interaction of acceleration rate and log(mpg) as explanatory variables. Specifically, acceleration rate and mile-per-gallon both positively affect the suggested retail price, but their interaction moderates the effect. This statistically confirms the common sense that hybrid car models with better acceleration performance are marketed at higher prices. We identify three representative brands affecting the suggested retail price. Brand Cadillac and Mercedes-Benz tend to have higher retail prices for their cars, but brand Jeep tends to have lower price. As discussed in the outlier test, since we only have one observation under the Jeep brand, we need more data to elaborate in future works. Cars with two seats have significantly lower price than other car classes. This agrees with intuition. Controlling for all other variables, the manufacturing cost of a two-seat car is lower than a regular four-seat car.

In application, future buyers can balance between the price-sensitive features above and other price-irrelevant features when shopping for hybrid electric vehicles, so that satisfaction can be maximized under given budget limits.

References

[1] Shabnam R.; Anderson Timothy R.; Tudorie Anca-Alexandra Lim, Dong-Joon; Jahromi. Comparing technological advancement of hybrid electric vehicles (hev) in different market segments. *Technological Forecasting Social Change*, 97:140–153, 08 2015.

[2] Homework assignment for sta 302 at university of toronto.

[3] Homework assignment for stat 425 at university of illinois at urbana-champaign.

[4] Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998.

[5] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 03 1978.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[7] David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.