

MSDS 6372 Project 1

SAMUEL ARELLANO and TRAVIS DAUN, Applied Statistics:Inference and Modeling, MSDS6372, Southern Methodist University

1 INTRODUCTION

Since Toyota's introduction of the Prius in 1997, there has been great interest in hybrid electric vehicles (HEVs) and a growing adoption of these vehicles by consumers. This has spurred other auto manufacturers to release their own versions of alternative fuel vehicles and led to a wide variety of HEVs for consumers to choose from. While the feature that consumers typically look for, and the one most marketed by manufacturers, is that of fuel efficiency measured in mile per gallon (MPG), one feature often overlooked is acceleration. Acceleration is not only an important performance metric, but like MPG, increases in acceleration rates of a vehicle class can be an indicator of technological advancement. In this study will review HEV data to determine what vehicle characteristics most influence a vehicle's acceleration rates, attempt to create a model based on the data that accurately predicts acceleration, and determine if acceleration rates across various classes of automobiles have improved over time.

2 DATA DESCRIPTION

The dataset is composed of 153 observations. Each observation is distinct model year HEV consisting of 9 variables (carid, vehicle, year, msrp, accelerate, mpg, mpgmpge, carclass, carclass_id). Each observation is identified by a unique carid. The HEV model is defined by the variable vehicle. Model year is defined by year. MSRP refers to the manufacturer's suggested retail price normalized to 2013 dollars. The acceleration rate is recorded as accelerate and represents the time in seconds required for the vehicle to reach a speed of 60 mph from a complete stop. Fuel economy is measured by the variable MPG. The variable MPGMPGe records the higher of the vehicles MPG or MPG equivalent estimated as $MPG_e = \frac{33.7 \times \text{driving range}}{\text{battery capacity}}$. The vehicles are categorized into 7 classes (C = Compact, M = Midsize, TS = 2 Seater, L = Large, PT = Pickup Truck, MV = Minivan, SUV = Sport Utility Vehicle) denoted under the variable carclass and each class is assigned a number denoted under the variable carclass_id.

Dataset: hybrid_reg.csv

Source: D-J. Lim, S.R. Jahromi, T.R. Anderson, A-A. Tudorie (2014). "Comparing Technological Advancement of Hybrid Electric Vehicles (HEV) in Different Market Segments," Technological Forecasting & Social Change, <http://dx.doi.org/10.1016/j.techfore.2014.05.008>

3 EXPLORATORY ANALYSIS

After an initial review of the data it was determined that the variable carid would not be used in linear modeling as its sole purpose was the identification of observations and because the vehicles were ordered by year and carid was assigned incrementally, using carid would invariably lead to collinearity problems with year without adding any information of value.

The categorical variables, vehicles and carclass_id were tested against accelerate to determine if the differences in mean acceleration rates were significantly different when grouped by

Authors' address: Samuel Arellano, sarellano@mail.smu.edu; Travis Daun, trdaun@mail.smu.edu, Applied Statistics:Inference and Modeling, MSDS6372, Southern Methodist University.

model or car class. We found that using an analysis of variance, the effect of vehicle was significant with a p-value of <.0001 and its R-Square value of .953 suggested that it might be a good explanatory variable, this can be seen in Figure 1 . Similarly, as seen in Figure 2 we found the effect of carclass_id was statistically significant with a p-value of <.0001, though at .337 it had a lower R-Squared value.

To test the fit of the continuous variables, we ran a scatterplot matrix and color coded the plots by car class. As can be seen in Figure 3, visual analysis of the scatterplot matrix indicated a normal distribution of the response variable accelerate and the presence of a good linear correlation between accelerate and msrp. It also indicated that there might be a linear correlation between accelerate and year if controlled for carclass. In Figure 4, we suspected that there might be a curvilinear relationship between accelerate and the variables mpg and mpgmpge, however after transforming the variables we decided that what we were seeing was the presence of outliers and not an indication of a nonlinear relationship. We performed a logarithmic transformation of mpg and mpgmpge to see if it would help improve linearity. We found that while the transformation made a marginal improvement on linear correlation and the logged values were more normally distributed, the improvement was not enough to warrant altering the data and since normal distribution of the predictor variables is not necessary for conducting a multiple regression, we opted to move forward with non-transformed data. This can be seen in Figure 5 and Figure 6. Analysis of the scatter plots also revealed the presence of collinearity between year and carid and between mpg and mpgmpge. The strong relation between carid and year was expected due to the data being sorted by year prior to the sequential assignment of car identification numbers. Due to the small quantity of all-electric HEV's, nearly all observations of mpg and mpgmpge were the same, resulting in strong collinearity between the two variables.

It was also important to understand the interactions between the various predictors. Figure 7 shows how the interaction between mpg and carclass impacts accelerate. We asked ourselves two main questions when exploring interactions; 1) Does the interaction make sense, and 2) Is the interaction statistically significant? For the interaction between carclass and mpg we determined that yes, some interaction would make sense, but as can be seen in Figure 7 the interaction is not statistically significant. We also explored the interaction between carclass and msrp which can be seen in Figure 8. Again intuition told us that this interaction makes conceptual sense since car prices are greatly impacted by the type of car you are buying. Again when the statistical significance of this interaction is explored, Figure 8 shows that the interaction between carclass and msrp are not statistically significant.

Because the value of carid has no real effect on the question of interest we decided to go ahead and remove it from further analysis. We then had to decide if we wanted to proceed with either mpg or mpgmpge in linear modeling. From the correlation matrix we found that the Pearson Correlation Coefficients of mpg and mpgmpge were -.506 and -.399 respectively, so we decided to explore mpg further as we moved into linear modeling. Before moving on we decided to run more correlation tests and explored various transformations of mpg against accelerate to ensure that our previous conclusion that the apparent curve in the line was nothing more than the presence of outliers. With the exploratory data analysis complete we moved on to linear modeling.

4 OBJECTIVE 1

4.1 Restatement of Problem and the overall approach to solve it

We wanted to determine if the acceleration rate of a HEV could be predicted by the price paid for the car, the miles per gallon that the car achieved, and the year that the car was manufactured. We chose to apply a multiple linear regression models using various sub set selection techniques. Specifically we employed stepwise, least angle regression (LAR), and least absolute shrinkage and selection operator (LASSO) algorithms for our feature selection. We also explored interactions between terms as well as exploring cross validation to fit the model.

4.2 Model Selection

4.2.1 *Checking Assumptions.* In order to use a multiple linear regression model the assumptions of linearity, multivariate normality, homoscedasticity, and no multicollinearity had to be met. To do a preliminary test of these assumptions we ran a regression model of the continuous variables, excluding carid, as seen in Figure 9 and looked at its resulting plots, fit diagnostics, variance inflation factors, and correlation matrix. Review of the scatter plots validated that a linear relationship existed, the quantile plot of the residuals indicated that multivariate normality existed, and plots of residuals against predicted values validated homoscedasticity. Though we had intended to drop the mpgmpge variable from the model due to the findings of our initial correlation analysis, when we tested a first regression model we decided to include both mpg and mpgmpge and we found that with variance inflation factors of 2.224 and 1.858, though correlation among the variables existed perhaps it was not to the degree we first thought. This was further validated by the correlation matrix which indicated that the correlation between the two was .668, a little high but under the .800 threshold for multicollinearity. Additionally, though the p-value for both was high .339 and .056 when both variables were included, it appeared that mpgmpge was the more significant of the two. Because of this we decided to not exclude mpgmpge from further linear modeling. Lastly we looked at studentized residual leverage plots and while we found the presence of outliers, none seem to have significant leverage. Furthermore, review of the Cook's D plot also indicated the presence of outliers, but since none where over .08 we decided to move forward onto model testing hoping to find the model that best predicted accelerate.

4.2.2 *Compare Competing Models.* The first model selection process used year msrp mpg mpgmpge and carclass_id as the predictors, no cross validation or interactions, and a stepwise selection method. The model returned used year, msrp, and mpgmpge to predict accelerate. It did so with an adjusted R-squared value of .552 and AICC of 366.474, we would use this as our benchmark for further testing. We repeated the selection process using the same variables but different selection methods. Both Least-Angle Regression (LAR) and Least Absolute Shrinkage and Selection Operator (LASSO) resulted in a model using msrp as the only predictor and yielded an adjusted R-squared value of .395 and AICC of 410.205, worse than the stepwise selection. We decided that in order to get a better model we would have to incorporate interaction variables and cross validation.

We chose the following set of interaction variables: msrp*year, msrp*carclass_id, year*carclass_id, year*vehicle, vehicle*carclass_id, and mpgmpge*carclass_id based on vehicle attributes known to be paired like model and model year and vehicle type and price. We ran the three model selection processes, but ended up with only marginally better results. The Stepwise AICC improved from 366.474 to 363.652 and the AICC for LAR and LASSO improved less

than a point from 261.105 to 260.734.

To see if we could decrease the AICC and increase the adjusted R-squared value even more we decided to try model selection with cross validation. To balance bias and variance we chose 5 fold random cross validation using a set seed and partitioned the data as: 50% Train, 25% Test, and 25% Validation. Stepwise selection returned a model with predictor variables and resulted in an improvement in the adjusted R-squared and AICC values .619 and 182.429 respectively. Use of cross validation also resulted in improved results for the LAR and LASSO selection methods, though like in previous tests these results were identical with both resulting in a 2 variable model with a 2.041 R-squared value and 199.653 AICC. We decided to try ELASTICNET as an additional selection method which yielded a 3 variable model that had a 2.055 adjusted R-squared value and a 201.986 AICC.

Method	Predictors	RMSE	Adj R²	AICC	SBC
Stepwise no Interaction no CV	3	1.968	0.552	366.474	223.187
Stepwise w/ Interaction no CV	3	1.950	0.569	363.652	220.366
Stepwise w/ Interaction and CV	3	1.819	0.619	182.429	109.146
LARS no Interaction no CV	1	2.286	0.395	410.205	261.105
LARS w/ Interaction no CV	1	2.284	0.397	409.673	260.734
LARS w/ Interaction and CV	2	2.041	0.520	199.653	124.266
LASSO no Interaction no CV	1	2.286	0.395	410.205	261.105
LASSO w/ Interaction no CV	1	2.284	0.397	409.673	260.734
LASSO w/ Interaction and CV	2	2.041	0.520	199.653	124.266
ELASTICENT no Interaction no CV	9	1.961	0.550	372.661	246.093
ELASTICENT w/ Interaction no CV	1	2.284	0.397	409.673	260.734
ELASTICENT w/ Interaction and CV	3	2.055	0.513	201.986	128.703

Table 1. Analysis Results

Though the adjusted R-squared value was lower than we would have liked, in the end the best model was a 3 variable model achieved through stepwise selection.

$$\text{accelerate} = 10.56 - 0.007\text{msrp} - 0.05\text{mpg} + 0.000003924\text{year}^*\text{msrp}$$

We then took this final model and performed a regression analysis to check it for fit as we had done with the initial model. We also ran partial F-test to validate that the interaction term made this model more statistically significant. The results of this partial F-test can be seen in Figure 10 and with a p-value <.001 concludes that the model with the interaction term is significantly better than without.

4.3 Parameter Interpretation

4.3.1 Interpretation.

4.3.2 Confidence Intervals.

4.4 Final conclusions from the analyses of Objective 1

5 OBJECTIVE 2

5.1 2way ANOVA

We want to see if carclass interacts with the price of a HEV when determining the acceleration rate of the vehicle. A categorical variable price having two factors, high and low as determined by the msrp of the HEV was developed. If the msrp is greater than the median msrp of all HEVs, price is set to high, else price is set to low.

5.2 Main Analysis Content

Figure 11 shows the plot of accelerate vs price grouped by carclass. From looking at this graph it is apparent that the lower priced cars in each category have lower acceleration as compared to higher priced cars within the same category. The graph also shows that higher priced cars in general have higher rates of acceleration. The differences in the accelerate between the high and low price may show an interaction effect since the clusters on the low side are much closer for each group than on the high side.

To test for the interaction effect a two-way ANOVA was run using an interaction between price and carclass. A linear model was used to model the interaction effect on accelerate between price and carclass. The ANOVA was then run on this model. The results can be seen in Table 2. From these results we can see that the interaction between price and carclass is highly statistically significant with a p-value of the interaction term of .001.

Anova Table (Type II tests)

Response: accelerate

	Sum Sq	Df	F value	Pr(>F)	
price	267.65	1	70.6647	4.175e-14	***
carclass	191.15	6	8.4112	8.349e-08	***
price:carclass	69.28	4	4.5727	0.001676	**
Residuals	534.06	141			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Table 2. Two Way ANOVA.

As with any ANOVA, we must meet the required assumptions of normal distribution of model residuals and homogeneity of variance of the groups. For the first assumption, normal distribution of model residuals, we generated a qq plot of the model residuals as seen in Figure 12. As can be seen in this figure, there is not an apparent departure from normality so we accept that this assumption holds. The second assumption also holds as seen in Figure 13 since the plot shows a random cloud that supports homogeneity of variance. As a validation of these assumptions a Shapiro-Wilk normality test and Levene's Test for Homogeneity of Variance were also performed which support these assumption are met. Table 3 shows these results. While the p-value of the Shapiro-Wilk normality test is inconclusive for normality, the QQ plot in Figure 12 gives us confidence of

Normality	Equal Variance
<p>Shapiro-Wilk normality test</p> <pre>data: aov_residuals W = 0.97532, p-value = 0.007489</pre>	<p>Levene's Test for Homogeneity of Variance (center = median)</p> <pre>Df F value Pr(>F) group 11 1.3613 0.1977 141</pre>

Table 3. Two Way ANOVA Assumptions.

normality. Since we do have a significant interaction effect, we want to look at all the separate group combinations. Figure 14 shows the results from this analysis. Table 4 shows a summary of the lsmeans between each carclass and each price combination.

price	carclass	lsmean	SE	df	lower.CL	upper.CL
high	C	12.36	1.124	141	10.14	14.58
low	C	9.58	0.361	141	8.86	10.29
high	L	16.49	0.736	141	15.03	17.94
low	L	12.35	1.946	141	8.50	16.20
high	M	16.64	0.540	141	15.57	17.71
low	M	11.42	0.297	141	10.83	12.00
high	MV	nonEst	NA	NA	NA	NA
low	MV	7.85	0.973	141	5.93	9.77
high	PT	11.18	1.376	141	8.45	13.90
low	PT	10.86	0.973	141	8.94	12.79
high	SUV	13.52	0.382	141	12.77	14.28
low	SUV	11.71	0.540	141	10.65	12.78
high	TS	nonEst	NA	NA	NA	NA
low	TS	9.85	0.688	141	8.49	11.21

Table 4. Posthoc Analysis of Interactions.

5.3 Conclusion/Discussion

Using the Two-Way ANOVA, we can look at the separate group combinations between price and carclass. Looking at Table 14 we can see which combinations of interactions are statistically significant. Using a 95% confidence interval and applying a tukey adjustment for comparing the family of 14 estimates we see that the following statistically significant interpretations can be made:

low, C - high, L Holding all things constant, a HEV in carclass C with a price less than \$40,000 will have an acceleration rate 6.9 seconds slower than a HEV in carclass L with a price greater than \$40,000. So a large HEV with an msrp greater than \$40,000 will reach 60 miles per hour 6.9 seconds faster than a compact HEV with a msrp less than \$40,000.

low, C - high, M Holding all things constant, a HEV in carclass C with a price less than \$40,000 will have an acceleration rate 7.0 seconds slower than a HEV in carclass M with a price greater than \$40,000. So a midsize HEV with an msrp greater than \$40,000 will reach 60 miles per hour 7.0 seconds faster than a compact HEV with a msrp less than \$40,000.

low, C - high, SUV Holding all things constant, a HEV in carclass C with a price less than \$40,000 will have an acceleration rate 3.9 seconds slower than a HEV in carclass SUV with a price greater than \$40,000. So a sport utility HEV with an msrp greater than \$40,000 will reach 60 miles per hour 3.9 seconds faster than a compact HEV with a msrp less than \$40,000.

high, L - low, M Holding all things constant, a HEV in carclass L with a price greater than \$40,000 will have an acceleration rate 5.1 seconds faster than a HEV in carclass M with a price less than \$40,000. So a large HEV with an msrp greater than \$40,000 will reach 60 miles per hour 5.1 seconds faster than a midsize HEV with a msrp less than \$40,000.

high, L - low, MV Holding all things constant, a HEV in carclass L with a price greater than \$40,000 will have an acceleration rate 8.6 seconds faster than a HEV in carclass MV with a price less than \$40,000. So a large HEV with an msrp greater than \$40,000 will reach 60 miles per hour 5.1 seconds faster than a mini van HEV with a msrp less than \$40,000.

high, L - low, TS Holding all things constant, a HEV in carclass L with a price greater than \$40,000 will have an acceleration rate 6.6 seconds faster than a HEV in carclass TS with a price less than \$40,000. So a large HEV with an msrp greater than \$40,000 will reach 60 miles per hour 6.6 seconds faster than a two-seater HEV with a msrp less than \$40,000.

high, M - low, M Holding all things constant, a HEV in carclass M with a price greater than \$40,000 will have an acceleration rate 5.2 seconds faster than a HEV in carclass M with a price less than \$40,000. So a midsize HEV with an msrp greater than \$40,000 will reach 60 miles per hour 5.2 seconds faster than a midsize HEV with a msrp less than \$40,000.

high, M - low, MV Holding all things constant, a HEV in carclass M with a price greater than \$40,000 will have an acceleration rate 8.8 seconds faster than a HEV in carclass MV with a price less than \$40,000. So a midsize HEV with an msrp greater than \$40,000 will reach 60 miles per hour 8.8 seconds faster than a mini van HEV with a msrp less than \$40,000.

high, M - low, SUV Holding all things constant, a HEV in carclass M with a price greater than \$40,000 will have an acceleration rate 4.9 seconds faster than a HEV in carclass SUV with a price less than \$40,000. So a midsize HEV with an msrp greater than \$40,000 will reach 60 miles per hour 4.9 seconds faster than a sport utility HEV with a msrp less than \$40,000.

high, M - low, TS Holding all things constant, a HEV in carclass M with a price greater than \$40,000 will have an acceleration rate 6.8 seconds faster than a HEV in carclass TS with a price less than \$40,000. So a midsize HEV with an msrp greater than \$40,000 will reach 60 miles per hour 6.8 seconds faster than a two-seater HEV with a msrp less than \$40,000.

low, MV - high, SUV Holding all things constant, a HEV in carclass MV with a price less than \$40,000 will have an acceleration rate 5.7 seconds slower than a HEV in carclass SUV with a price greater than \$40,000. So a sport utility HEV with an msrp greater than \$40,000 will reach 60 miles per hour 5.7 seconds faster than a mini van HEV with a msrp less than \$40,000.

These results are very beneficial in determining which vehicle is most appropriate for the type of driving that is desired. From these results it is easy to see that if the acceleration rate is important to a consumer, then in general vehicles with an msrp greater than \$40,000 will have higher acceleration rates. It was disappointing that none of the group combinations of identical carclass were not statistically significant, nor were group combinations of identical price combinations. The two-way ANOVA provided insight into some of the factors that were not apparent in the multiple linear regression model.

A GRAPHICS AND SUMMARY TABLES

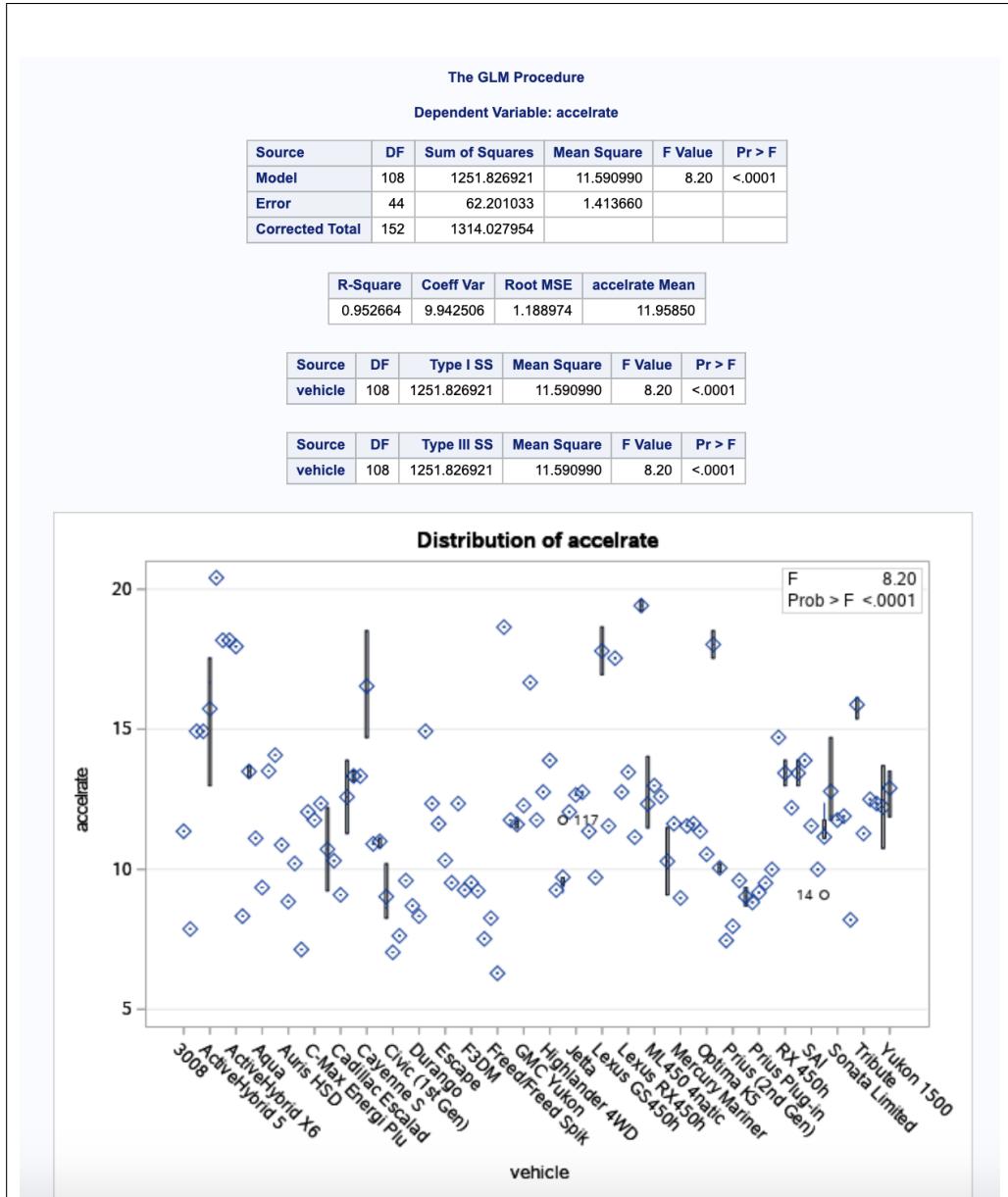


Fig. 1. Accelerate by Vehicle.

The GLM Procedure

Dependent Variable: accelerate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	443.032760	73.838793	12.38	<.0001
Error	146	870.995194	5.965721		
Corrected Total	152	1314.027954			

R-Square	Coeff Var	Root MSE	accelerate Mean
0.337156	20.42466	2.442482	11.95850

Source	DF	Type I SS	Mean Square	F Value	Pr > F
carclass_id	6	443.0327601	73.8387933	12.38	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
carclass_id	6	443.0327601	73.8387933	12.38	<.0001

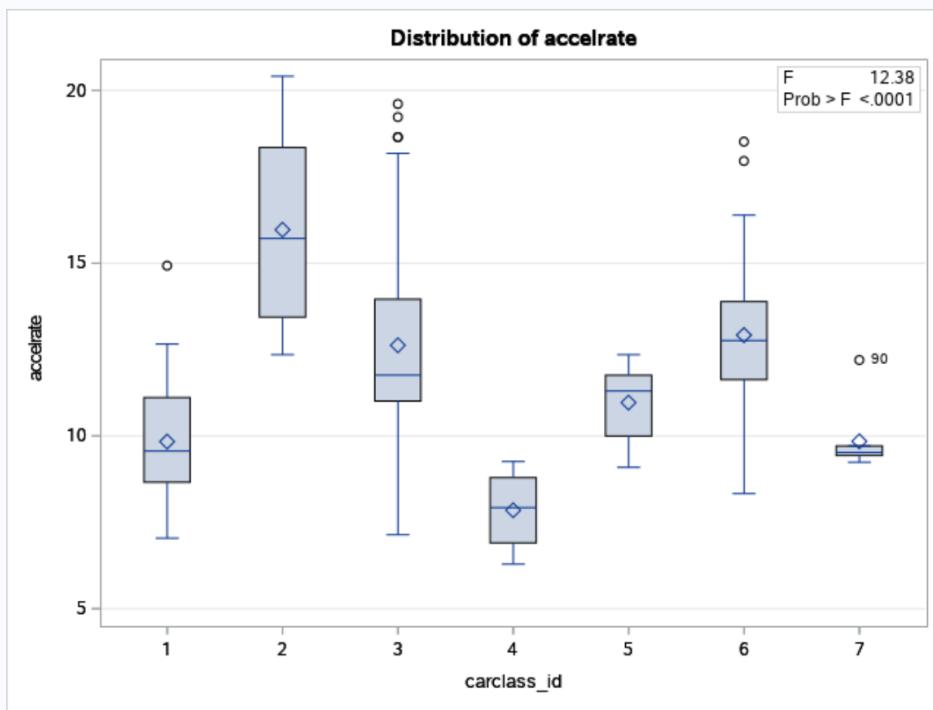


Fig. 2. Accelerate by Car Class.

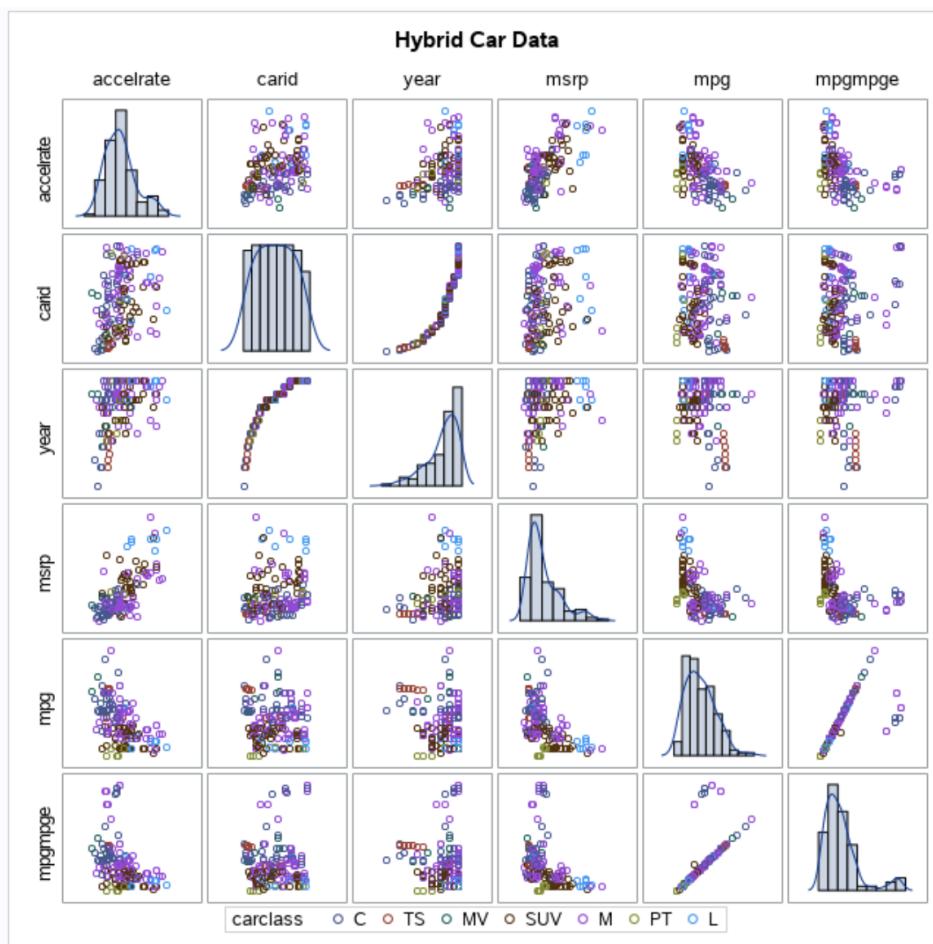


Fig. 3. Scatter Plot Matrix by Car Class.

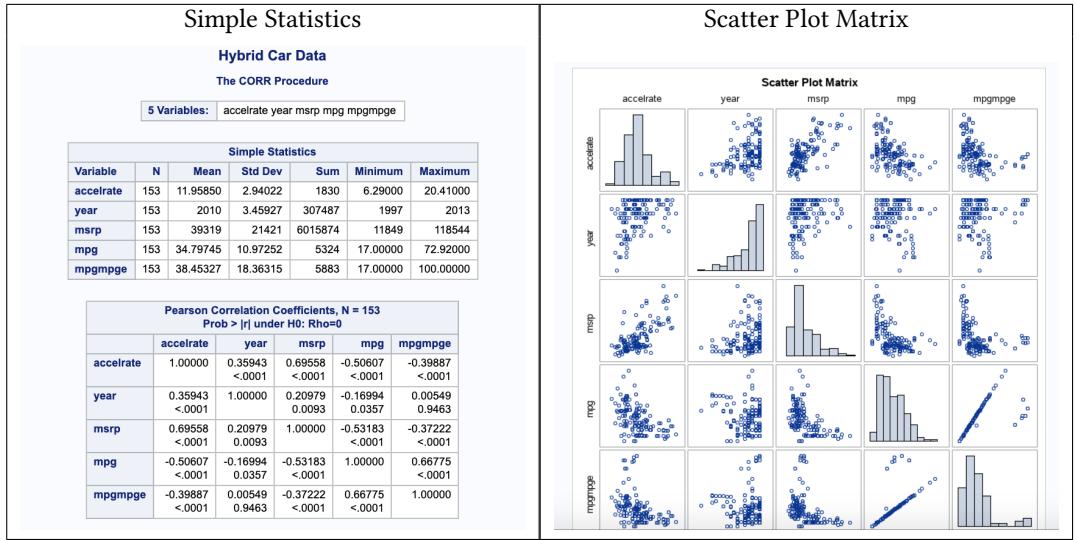


Fig. 4. Hybrid Car Data Correlation.

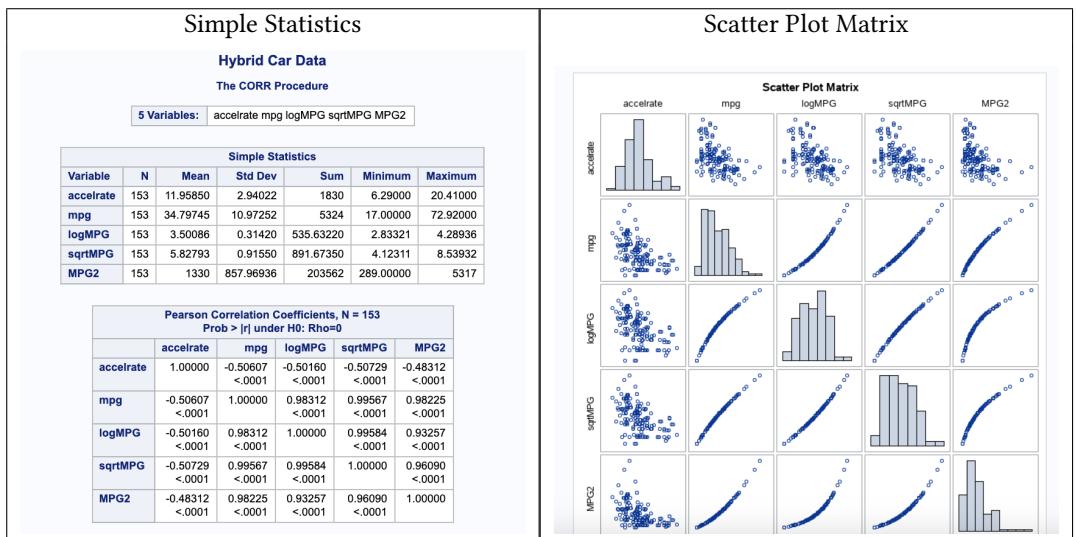


Fig. 5. Transformations of MPG vs accelerate.

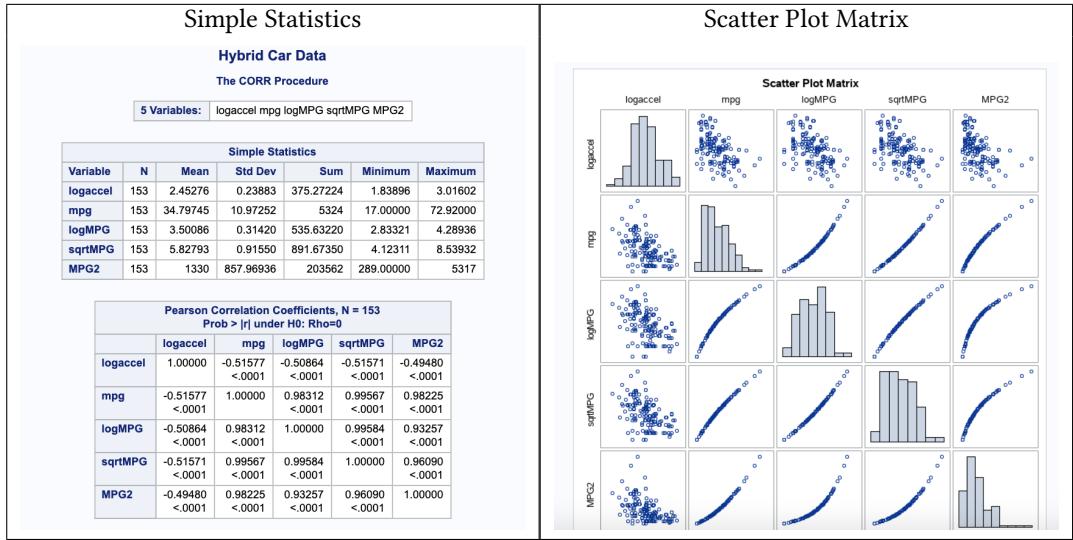


Fig. 6. Transformations of MPG vs logaccelerate.

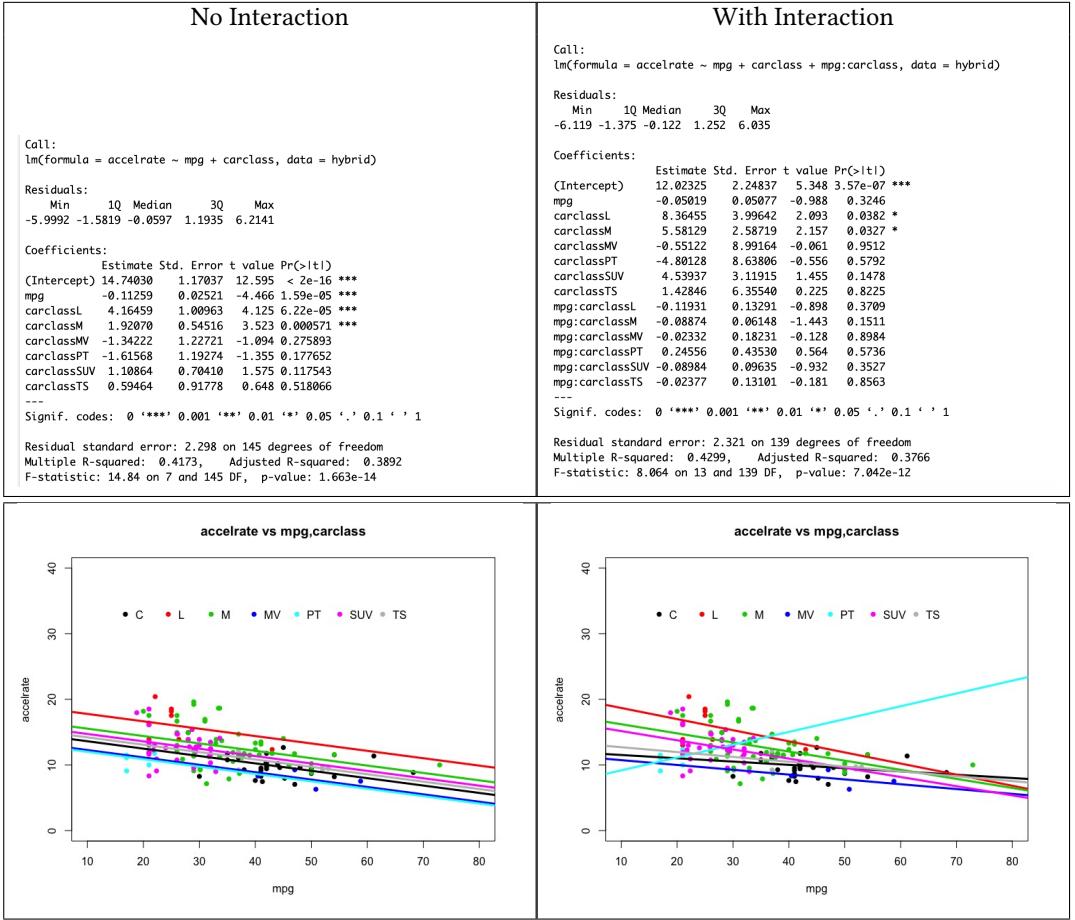


Fig. 7. Interactions between MPG and Car Class on Accelerate.

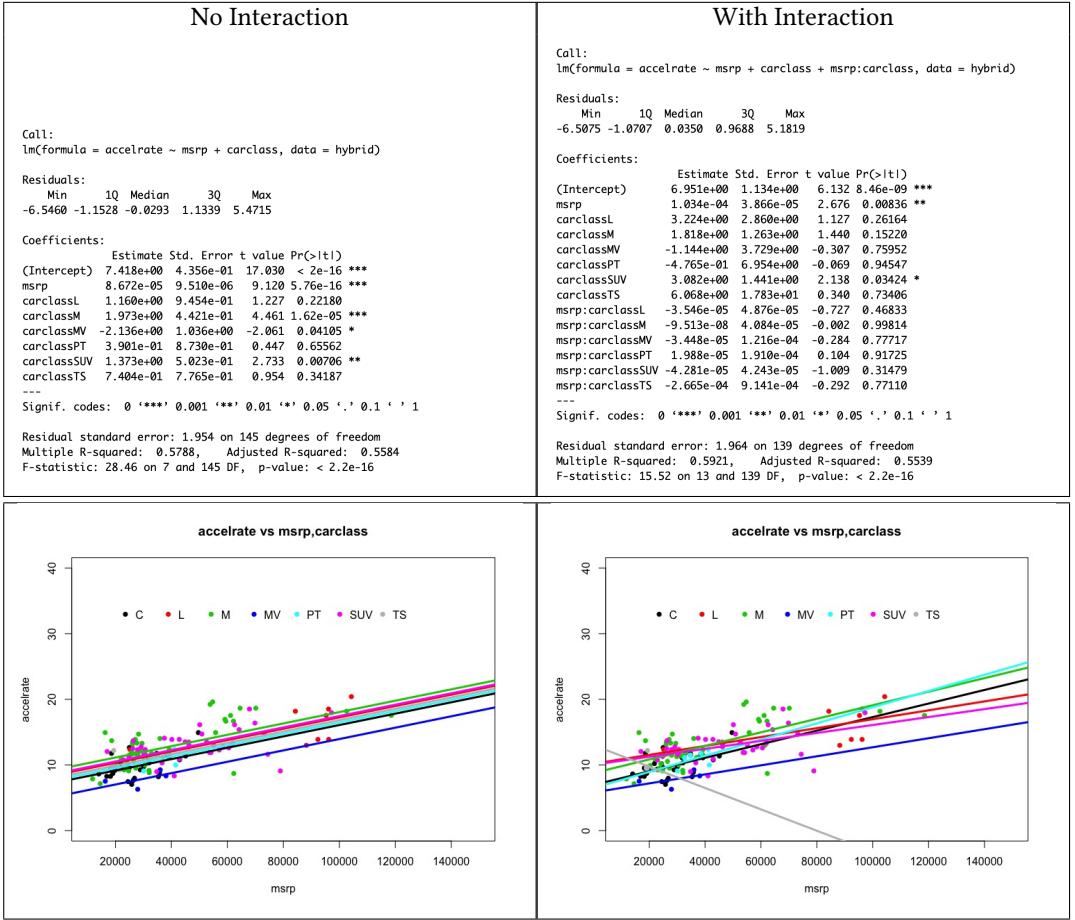


Fig. 8. Interactions between MSRP and Car Class on Accelerate.

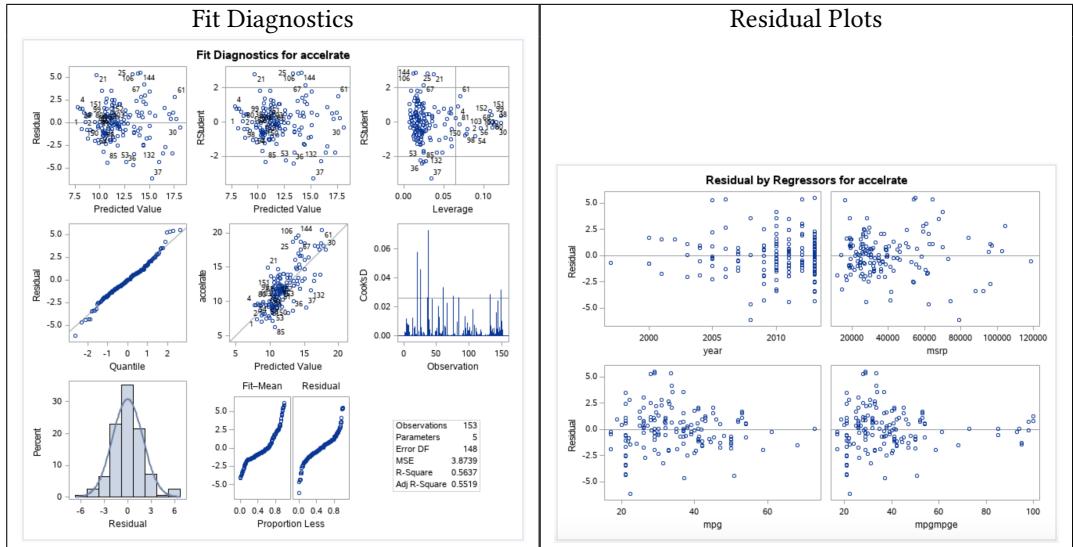


Fig. 9. Assumption Diagnostics for Hybrid Dataset.

```
> anova(Reduced.model, Full.model)
Analysis of Variance Table

Model 1: accelerate ~ msrp + mpg
Model 2: accelerate ~ msrp + mpg + year:msrp
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     150 644.30
2     149 576.32  1    67.986 17.577 4.713e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 10. Partial F-Test for Final Model.

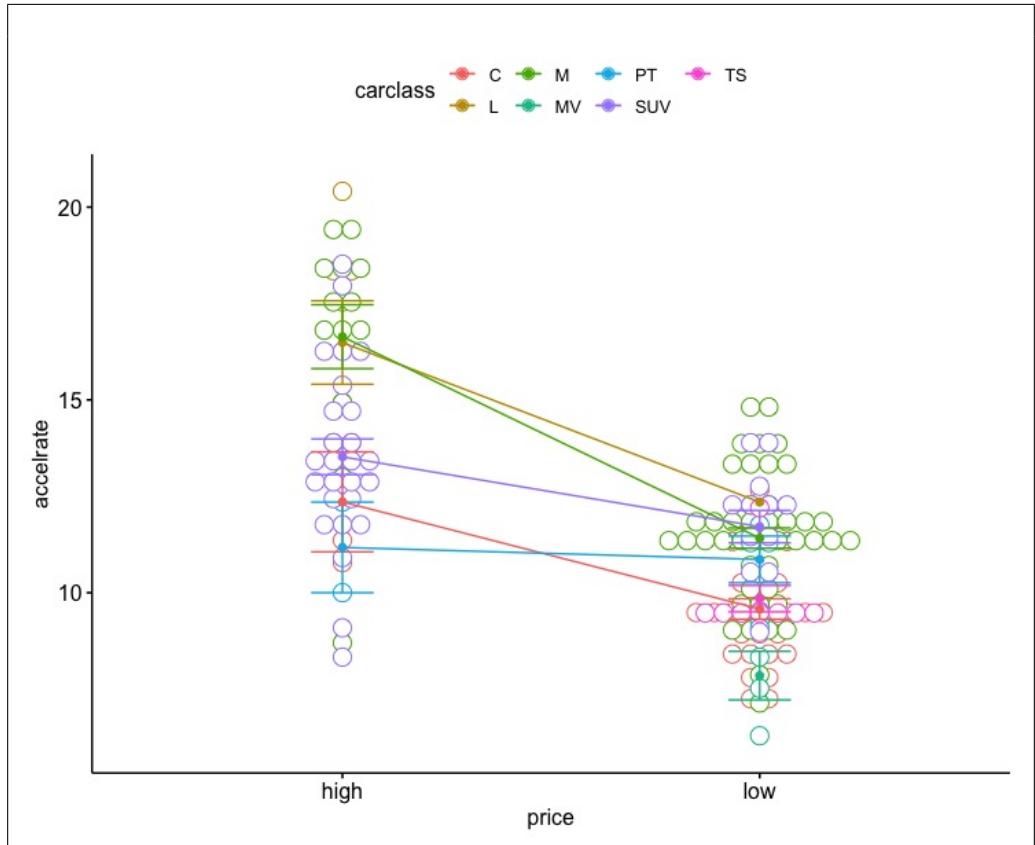


Fig. 11. Plot of Accelerate vs Price by Car Class.

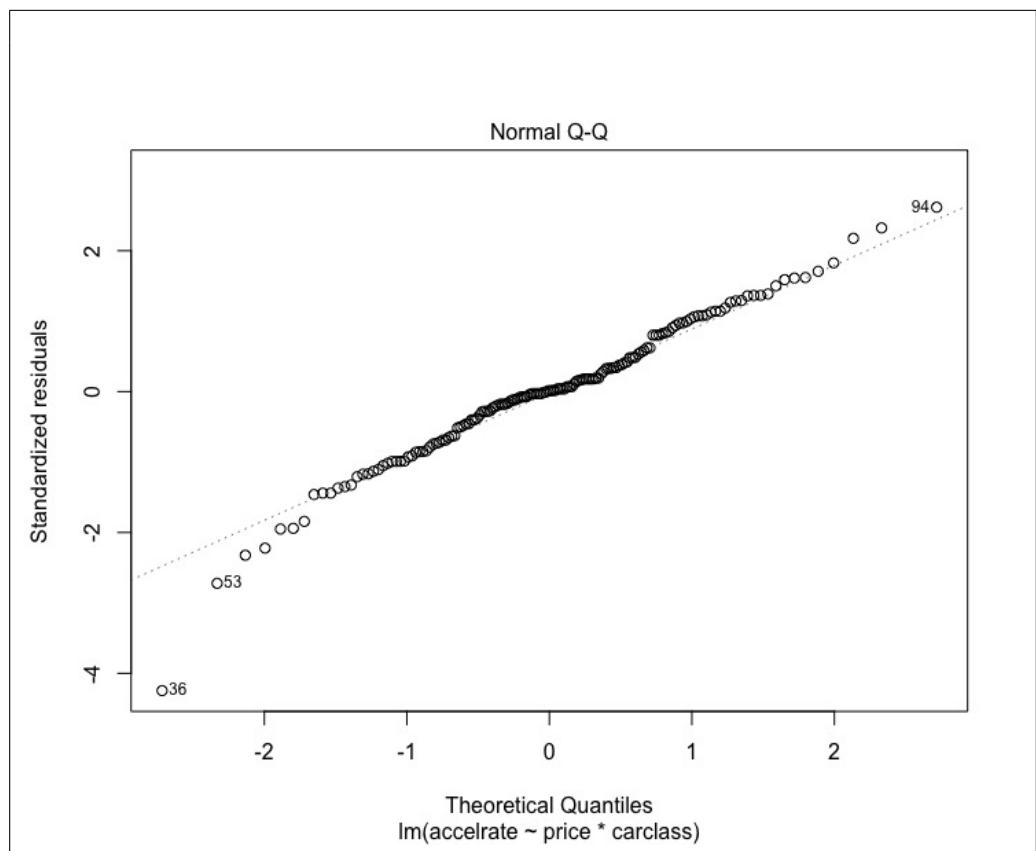


Fig. 12. QQ Plot of Model Residuals.

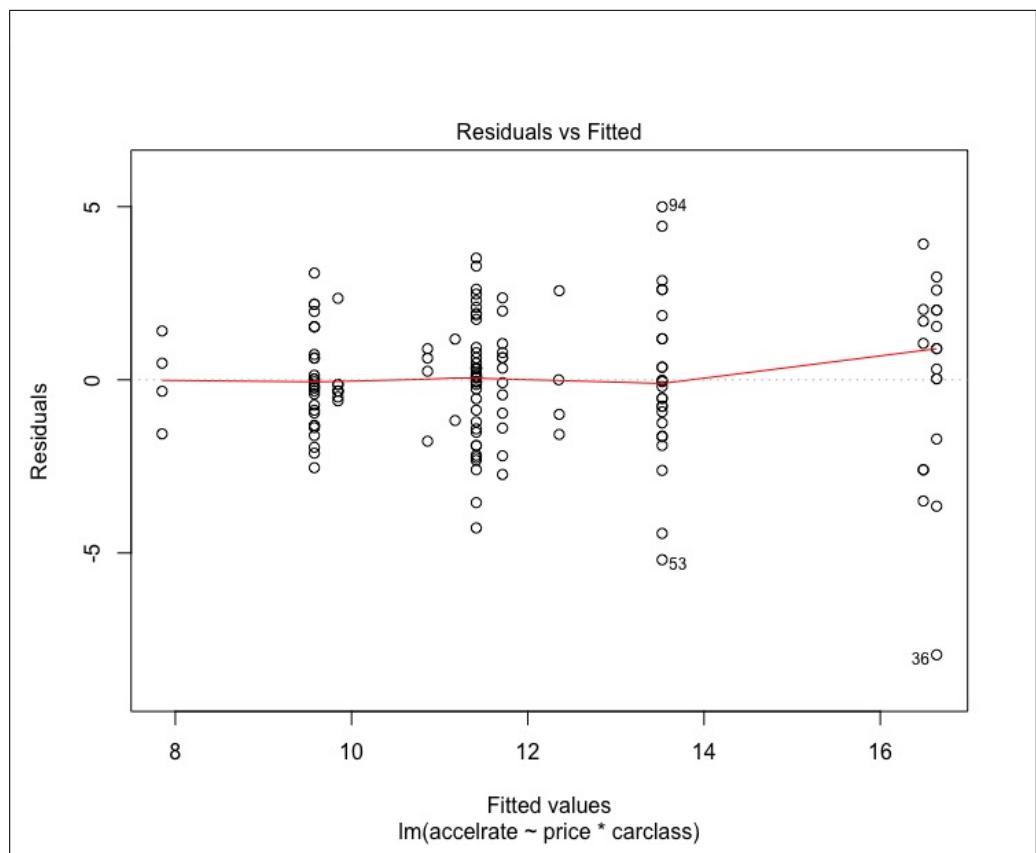


Fig. 13. Fitted Model Residuals.

\$contrasts	estimate	SE	df	t.ratio	p.value
contrast					
high,C - low,C	2.77908	1.180	141	2.355	0.5183
high,C - high,L	-4.13190	1.343	141	-3.077	0.1250
high,C - low,L	0.08667	2.247	141	0.003	1.0000
high,C - high,M	-4.28256	1.247	141	-3.436	0.0468
high,C - low,M	0.94108	1.162	141	0.810	0.9999
high,C - high,MV	nonEst	NA	NA	NA	NA
high,C - low,MV	4.58667	1.486	141	3.032	0.1396
high,C - high,PT	1.18167	1.777	141	0.665	1.0000
high,C - low,PT	1.49417	1.486	141	1.005	0.9992
high,C - high,SUV	-1.16795	1.187	141	-0.984	0.9993
high,C - low,SUV	0.64359	1.247	141	0.516	1.0000
high,C - high,TS	nonEst	NA	NA	NA	NA
high,C - low,TS	2.51042	1.318	141	1.905	0.8202
low,C - high,L	-6.91099	0.820	141	-8.432	<.0001
low,C - low,L	-2.77241	1.979	141	-1.401	0.9805
low,C - high,M	-7.06165	0.650	141	-10.871	<.0001
low,C - low,M	-1.83808	0.468	141	-3.938	0.0095
low,C - high,MV	nonEst	NA	NA	NA	NA
low,C - low,MV	1.72759	1.038	141	1.664	0.9247
low,C - high,PT	-1.59741	1.423	141	-1.123	0.9975
low,C - low,PT	-1.28491	1.038	141	-1.238	0.9935
low,C - high,SUV	-3.94703	0.526	141	-7.509	<.0001
low,C - low,SUV	-2.13549	0.650	141	-3.287	0.0716
low,C - high,TS	nonEst	NA	NA	NA	NA
low,C - low,TS	-0.26866	0.777	141	-0.346	1.0000
high,L - low,L	4.13857	2.081	141	1.989	0.7717
high,L - high,M	-0.15066	0.912	141	-0.165	1.0000
high,L - low,M	5.07299	0.793	141	6.396	<.0001
high,L - high,MV	nonEst	NA	NA	NA	NA
high,L - low,MV	8.63857	1.220	141	7.082	<.0001
high,L - high,PT	5.31357	1.560	141	3.405	0.0512
high,L - low,PT	5.62607	1.220	141	4.612	0.0007
high,L - high,SUV	2.96398	0.829	141	3.577	0.0305
high,L - low,SUV	4.77550	0.912	141	5.234	0.0001
high,L - high,TS	nonEst	NA	NA	NA	NA
high,L - low,TS	6.64232	1.007	141	6.595	<.0001
low,L - high,M	-4.28923	2.020	141	-2.124	0.6838
low,L - low,M	0.93442	1.969	141	0.475	1.0000
low,L - high,MV	nonEst	NA	NA	NA	NA
low,L - low,MV	4.50000	2.176	141	2.068	0.7214
low,L - high,PT	1.17500	2.384	141	0.493	1.0000
low,L - low,PT	1.48750	2.176	141	0.684	1.0000
low,L - high,SUV	-1.17461	1.983	141	-0.592	1.0000
low,L - low,SUV	0.63692	2.020	141	0.315	1.0000
low,L - high,TS	nonEst	NA	NA	NA	NA
low,L - low,TS	2.50375	2.061	141	1.213	0.9947
high,M - low,M	5.22365	0.616	141	8.480	<.0001
high,M - high,MV	nonEst	NA	NA	NA	NA
high,M - low,MV	8.78923	1.113	141	7.898	<.0001
high,M - high,PT	5.46423	1.478	141	3.696	0.0208
high,M - low,PT	5.77673	1.113	141	5.191	0.0001
high,M - high,SUV	3.11461	0.661	141	4.711	0.0005
high,M - low,SUV	4.92615	0.761	141	6.453	<.0001
high,M - high,TS	nonEst	NA	NA	NA	NA
high,M - low,TS	6.79298	0.875	141	7.768	<.0001
low,M - high,MV	nonEst	NA	NA	NA	NA
low,M - low,MV	3.56558	1.017	141	3.505	0.0380
low,M - high,PT	0.24058	1.408	141	0.171	1.0000
low,M - low,PT	0.55308	1.017	141	0.544	1.0000
low,M - high,SUV	-2.10903	0.483	141	-4.362	0.0019
low,M - low,SUV	-0.29750	0.616	141	-0.483	1.0000
low,M - high,TS	nonEst	NA	NA	NA	NA
low,M - low,TS	1.56933	0.749	141	2.094	0.7039
high,MV - low,MV	nonEst	NA	NA	NA	NA
high,MV - high,PT	nonEst	NA	NA	NA	NA
high,MV - low,PT	nonEst	NA	NA	NA	NA
high,MV - high,SUV	nonEst	NA	NA	NA	NA
high,MV - low,SUV	nonEst	NA	NA	NA	NA
high,MV - high,TS	nonEst	NA	NA	NA	NA
high,MV - low,TS	nonEst	NA	NA	NA	NA
low,MV - high,PT	-3.32500	1.685	141	-1.973	0.7816
low,MV - low,PT	-3.01250	1.376	141	-2.189	0.6379
low,MV - high,SUV	-5.67462	1.045	141	-5.429	<.0001
low,MV - low,SUV	-3.86308	1.113	141	-3.472	0.0421
low,MV - high,TS	nonEst	NA	NA	NA	NA
low,MV - low,TS	-1.99625	1.192	141	-1.675	0.9212
high,PT - low,PT	0.31250	1.685	141	0.185	1.0000
high,PT - high,SUV	-2.34962	1.428	141	-1.645	0.9306
high,PT - low,SUV	-0.53808	1.478	141	-0.364	1.0000
high,PT - high,TS	nonEst	NA	NA	NA	NA
high,PT - low,TS	1.32875	1.539	141	0.864	0.9998
low,PT - high,SUV	-2.66212	1.045	141	-2.547	0.3844
low,PT - low,SUV	-0.85058	1.113	141	-0.764	1.0000
low,PT - high,TS	nonEst	NA	NA	NA	NA
low,PT - low,TS	1.01625	1.192	141	0.853	0.9999
high,SUV - low,SUV	1.81154	0.661	141	2.740	0.2679
high,SUV - high,TS	nonEst	NA	NA	NA	NA
high,SUV - low,TS	3.67836	0.787	141	4.675	0.0006
low,SUV - high,TS	nonEst	NA	NA	NA	NA
low,SUV - low,TS	1.86683	0.875	141	2.135	0.6762
high,TS - low,TS	nonEst	NA	NA	NA	NA

P value adjustment: tukey method for comparing a family of 14 estimates

Fig. 14. Posthoc Test of Interactions.

B SAS CODE

Listing 1. SAS Code (AppliedStatsProject.sas).

```
1 proc import datafile="/home/sarellano0/dataSets/hybrid_reg.csv"
2         dbms=dlm out=Hybrid replace;
3         delimiter=' , ' ;
4         getnames=yes;
5
6 run;
7
8 /* accelrate ~ carid vehicle year msrp mpg mpgmpge carclass carclass_id */
9
10 /* bunch of data transformations */
11 data Hybrid2;
12 set Hybrid;
13 logaccel = log(accelrate);
14 logMPG = log(mpg);
15 logYear = log(year);
16 GP100 = 100/mpg;
17 logGP100 = log(100/mpg);
18 logMSRP = log(msrp);
19 sqrtMPG = sqrt(mpg);
20 sqrtMSRP = sqrt(msrp);
21 MPG2 = (mpg*mpg);
22 MSRP2 = (msrp*msrp);
23 run;
24
25
26 /* EDA of Statistical Significance of year as Categorical Variable */
27 proc GLM data=Hybrid2;
28     class year;
29     model accelrate=year;
30     run;
31
32 /* EDA of Statistical Significance of carclass_id as Categorical Variable */
33 proc GLM data=Hybrid2;
34     class carclass_id;
35     model accelrate=carclass_id;
36     run;
37
38 /* EDA of Continuous Variables */
39 proc sgscatter data=Hybrid2;
40     title "Hybrid Car Data";
41     matrix accelrate carid year msrp mpg mpgmpge / group=carclass diagonal=(0
42 run;
43     title;
```

```

45 /* EDA Correlation of Continous Variables */
46 ods graphics on;
47 title 'Hybrid Car Data';
48 proc corr data=Hybrid2 nomiss plots=matrix(histogram);
49   var accelerate year msrp mpg mpgmpge;
50   run;
51 ods graphics off;
52
53 /* EDA Correlation of MGP Transformations */
54 ods graphics on;
55 title 'Hybrid Car Data';
56 proc corr data=Hybrid2 nomiss plots=matrix(histogram);
57   var accelerate mpg logmpg sqrtMPG MPG2 ;
58   run;
59 ods graphics off;
60
61 /* EDA Correlation of MGP Transformations on logaccel */
62 ods graphics on;
63 title 'Hybrid Car Data';
64 proc corr data=Hybrid2 nomiss plots=matrix(histogram);
65   var logaccel mpg logmpg sqrtMPG MPG2 ;
66   run;
67 ods graphics off;
68
69 /* EDA examining data distribution of mpg */
70 proc univariate data = Hybrid2;
71 var mpg;
72 histogram;
73 run;
74
75 /* Linear Modeling Assumptions Continous Variables with VIF */
76 proc reg data = Hybrid2 corr plots(label) = all;
77   model accelerate = year msrp mpg mpgmpge/ VIF /*CLB*/ ;
78   title 'Hybrid Car Data';
79   run; quit;
80
81 /* Linear Modeling Assumptions All Variables */
82 proc glm data = Hybrid2 plots=all;
83   class carclass_id;
84   model accelerate = year msrp mpg mpgmpge carclass_id;
85 run;
86 quit;
87
88 /* Model Selection simple glmselect with no interaction variables Stepwise */
89 proc glmselect data = Hybrid2;
90   class carclass_id vehicle;
91   model accelerate = vehicle year msrp mpg mpgmpge carclass_id / sele
92   run; quit; /* year msrp mpgmpge

```

```

93                                         RMSE 1.968 Adj R-Sq      0.552 AICC | 366.474
94
95 /* Model Selection simple glmselect with no interaction variables LARS no
96 proc glmselect data = Hybrid2;
97     class carclass_id vehicle;
98     model accelerate = vehicle year msrp mpg mpgmpge carclass_id / selec
99     run; quit; /* msrp
100                                         RMSE      2.286 Adj R-Sq  0.395 AICC | 410.205
101
102 /* Model Selection simple glmselect with no interaction variables LASSO no
103 proc glmselect data = Hybrid2;
104     class carclass_id vehicle;
105     model accelerate = year msrp mpg mpgmpge carclass_id / selection =L
106     run; quit; /* msrp
107                                         RMSE      2.286 Adj R-Sq  0.395 AICC | 410.205
108
109 /* Model Selection simple glmselect with no interaction variables ELASTICN
110 proc glmselect data = Hybrid2;
111     class carclass_id vehicle;
112     model accelerate = year msrp mpg mpgmpge carclass_id / selection =E
113     run; quit; /* year msrp mpg mpgmpge carclass_id_1 carclass_id_2 ca
114                                         RMSE 1.961 Adj R-Sq      .550 AICC | 372.661
115
116
117 /* Model Selection simple glmselect with Interaction variables Stepwise no
118 proc glmselect data = Hybrid2;
119     class carclass_id vehicle;
120     model accelerate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
121     selection =STEPWISE;
122     run; quit; /* msrp mpgmpge year*msrp
123                                         RMSE 1.950 Adj R-Sq  0.569 AICC | 363.652 SBC
124
125 /* Model Selection simple glmselect with Interaction variables LAR no CV*/
126 proc glmselect data = Hybrid2;
127     class carclass_id vehicle;
128     model accelerate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
129     selection =LAR;
130     run; quit; /* year*msrp
131                                         RMSE 2.284 Adj R-Sq      0.397 AICC | 409.673
132
133 /* Model Selection simple glmselect with no interaction variables LASSO no
134 proc glmselect data = Hybrid2;
135     class carclass_id vehicle;
136     model accelerate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
137     selection =LASSO;
138     run; quit; /* year*msrp
139                                         RMSE 2.284 Adj R-Sq      0.397 AICC | 409.673
140

```

```

141 /* Model Selection simple glmselect with no interaction variables ELASTICNET */
142 proc glmselect data = Hybrid2;
143   class carclass_id vehicle;
144   model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*year;
145   selection =ELASTICNET;
146   run; quit; /* year*msrp
147               RMSE 2.284 Adj R-Sq      0.397 AICC 409.673 */
148
149
150 /* Model Selection glmselect with Interaction variables Stepwise with CV*/
151 proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A);
152   partition fraction(test = .25 validate = .25);
153   class carclass_id vehicle;
154   model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*year;
155   selection = stepwise( stop = CV) cvdetails;
156   run; quit; /* msrp mpgmpge year*msrp
157               RMSE 1.819 Adj R-Sq 0.619 AICC 182.429 SBC */
158
159 /* Model Selection glmselect with Interaction variables LAR with CV*/
160 proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A);
161   partition fraction(test = .25 validate = .25);
162   class carclass_id vehicle;
163   model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*year;
164   selection = LAR( stop = CV) cvdetails;
165   run; quit; /* mpg year*msrp
166               RMSE 2.041 Adj R-Sq 0.520 AICC 199.653 SBC */
167
168 /* Model Selection glmselect with Interaction variables LASSO with CV*/
169 proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A);
170   partition fraction(test = .25 validate = .25);
171   class carclass_id vehicle;
172   model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*year;
173   selection = LASSO( stop = CV) cvdetails;
174   run; quit; /* mpg year*msrp
175               RMSE 2.041 Adj R-Sq 0.520 AICC 199.653 SBC */
176
177 /* Model Selection glmselect with Interaction variables ELASTICNET with CV*/
178 proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A);
179   partition fraction(test = .25 validate = .25);
180   class carclass_id vehicle;
181   model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*year;
182   selection = ELASTICNET( stop = CV) cvdetails;
183   run; quit; /* msrp mpg year*msrp
184               RMSE 2.055 Adj R-Sq 0.513 AICC 201.986 SBC */
185
186 /* Final Model Fit Diagnostics */
187 proc glm data = Hybrid2 plots=all;
188   model accelrate = msrp mpg year*msrp;

```

189 | run ;
190 | quit ;

C R CODE

Listing 2. R Code (AppliedStatsProject.R).

```

1 #Load dataset
2 hybrid<-read.csv( "hybrid_reg.csv" )
3 summary(hybrid)
4 ## accelerate with mpg and carclass - no interaction
5 mod.mpg.class.nointer = lm(accelerate ~ mpg + carclass , data=hybrid)
6 par(mfrow=c(2,2))
7 plot(mod.mpg.class.nointer)
8 summary(mod.mpg.class.nointer)
9 par(mfrow=c(1,1))
10 plot(mpg[carclass=="C"], accelerate[carclass=="C"], col=1, pch=16,
11       xlab="mpg", ylab="accelerate",
12       main="accelerate vs mpg, carclass",
13       ylim=c(0,40), xlim=c(10,80))
14 points(mpg[carclass=="L"], accelerate[carclass=="L"],
15         col=2,pch=16)
16 points(mpg[carclass=="M"], accelerate[carclass=="M"],
17         col=3,pch=16)
18 points(mpg[carclass=="MV"], accelerate[carclass=="MV"],
19         col=4,pch=16)
20 points(mpg[carclass=="PT"], accelerate[carclass=="PT"],
21         col=5,pch=16)
22 points(mpg[carclass=="SUV"], accelerate[carclass=="SUV"],
23         col=6,pch=16)
24 points(mpg[carclass=="TS"], accelerate[carclass=="TS"],
25         col=8,pch=16)
26 abline(a=14.74030, b=-0.11259, col=1, lwd=3) #C
27 abline(a=(14.74030+4.16459), b=-0.11259,
28         col=2, lwd=3) #L
29 abline(a=(14.74030+1.92070), b=-0.11259,
30         col=3, lwd=3) #M
31 abline(a=(14.74030-1.34222), b=-0.11259,
32         col=4, lwd=3) #MV
33 abline(a=(14.74030-1.61568), b=-0.11259,
34         col=5, lwd=3) #PT
35 abline(a=(14.74030+1.10864), b=-0.11259,
36         col=6, lwd=3) #SUV
37 abline(a=(14.74030+0.59464), b=-0.11259,
38         col=8, lwd=3) #TS
39 legend(15,35, legend=
40         c("C", "L", "M", "MV", "PT", "SUV", "TS"),
41         col=c(1,2,3,4,5,6,8),
42         pch=c(16,16,16,16,16,16,16),
43         bty = "n", horiz=TRUE)
44

```

```

45 ## accelerate with mpg and carclass - interaction
46 mod.mpg.class.inter = lm(accelrate ~ mpg + carclass +
47                           mpg:carclass,
48                           data=hybrid)
49 par(mfrow=c(2,2))
50 plot(mod.mpg.class.inter)
51 summary(mod.mpg.class.inter)
52 par(mfrow=c(1,1))
53 plot(mpg[carclass=="C"], accelrate[carclass=="C"],
54       col=1, pch=16,
55       xlab="mpg", ylab="accelrate",
56       main="accelrate vs mpg, carclass",
57       ylim=c(0,40), xlim=c(10,80))
58 points(mpg[carclass=="L"], accelrate[carclass=="L"],
59         col=2,pch=16)
60 points(mpg[carclass=="M"], accelrate[carclass=="M"],
61         col=3,pch=16)
62 points(mpg[carclass=="MV"], accelrate[carclass=="MV"],
63         col=4,pch=16)
64 points(mpg[carclass=="PT"], accelrate[carclass=="PT"],
65         col=5,pch=16)
66 points(mpg[carclass=="SUV"], accelrate[carclass=="SUV"],
67         col=6,pch=16)
68 points(mpg[carclass=="TS"], accelrate[carclass=="TS"],
69         col=8,pch=16)
70 abline(a=12.02325, b=-0.05019, col=1, lwd=3) #C
71 abline(a=(12.023258 + 8.36455), b=(-0.05019 - 0.11931),
72         col=2, lwd=3) #L
73 abline(a=(12.02325+5.58129), b=(-0.05019 - 0.08874),
74         col=3, lwd=3) #M
75 abline(a=(12.02325-0.55122), b=(-0.05019 - 0.02332),
76         col=4, lwd=3) #MV
77 abline(a=(12.02325-4.80128), b=(-0.05019 + 0.24556),
78         col=5, lwd=3) #PT
79 abline(a=(12.02325+4.53937), b=(-0.05019 - 0.08984),
80         col=6, lwd=3) #SUV
81 abline(a=(12.02325+1.42846), b=(-0.05019 - 0.02377),
82         col=8, lwd=3) #TS
83 legend(15,35, legend=
84         c("C", "L", "M", "MV", "PT", "SUV", "TS"),
85         col=c(1,2,3,4,5,6,8), pch=c(16,16,16,16,16,16,16),
86         bty = "n", horiz=TRUE)
87
88 ## accelerate with msrp and carclass - no interaction
89 mod.msrp.class.nointer = lm(accelrate ~ msrp + carclass,
90                             data=hybrid)
91 par(mfrow=c(2,2))
92 plot(mod.msrp.class.nointer)

```

```

93 summary(mod.msdp.class.nointer)
94 par(mfrow=c(1,1))
95 plot(msdp[carclass=="C"], accelrate[carclass=="C"],
96   col=1, pch=16,
97   xlab="msdp", ylab="accelrate",
98   main="accelrate vs msdp, carclass",
99   ylim=c(0,40), xlim=c(10000,150000))
100 points(msdp[carclass=="L"], accelrate[carclass=="L"],
101   col=2,pch=16)
102 points(msdp[carclass=="M"], accelrate[carclass=="M"],
103   col=3,pch=16)
104 points(msdp[carclass=="MV"], accelrate[carclass=="MV"],
105   col=4,pch=16)
106 points(msdp[carclass=="PT"], accelrate[carclass=="PT"],
107   col=5,pch=16)
108 points(msdp[carclass=="SUV"], accelrate[carclass=="SUV"],
109   col=6,pch=16)
110 points(msdp[carclass=="TS"], accelrate[carclass=="TS"],
111   col=8,pch=16)
112 abline(a=7.418, b=8.672e-05, col=1, lwd=3) #C
113 abline(a=(7.418+1.160), b=8.672e-05, col=2, lwd=3) #L
114 abline(a=(7.418+1.973), b=8.672e-05, col=3, lwd=3) #M
115 abline(a=(7.418-2.136), b=8.672e-05, col=4, lwd=3) #MV
116 abline(a=(7.418+3.901e-01), b=8.672e-05, col=5, lwd=3) #PT
117 abline(a=(7.418+1.373), b=8.672e-05, col=6, lwd=3) #SUV
118 abline(a=(7.418+7.404e-01), b=8.672e-05, col=8, lwd=3) #TS
119 legend(20000,35, legend=
120   c("C", "L", "M", "MV", "PT", "SUV", "TS"),
121   col=c(1,2,3,4,5,6,8), pch=c(16,16,16,16,16,16,16),
122   bty = "n", horiz=TRUE)
123
124 ## accelrate with msdp and carclass - interaction
125 mod.msdp.class.inter = lm(accelrate ~ msdp + carclass +
126                           msdp:carclass,
127                           data=hybrid)
128 par(mfrow=c(2,2))
129 plot(mod.msdp.class.inter)
130 summary(mod.msdp.class.inter)
131 par(mfrow=c(1,1))
132 plot(msdp[carclass=="C"], accelrate[carclass=="C"],
133   col=1, pch=16,
134   xlab="msdp", ylab="accelrate",
135   main="accelrate vs msdp, carclass",
136   ylim=c(0,40), xlim=c(10000,150000))
137 points(msdp[carclass=="L"], accelrate[carclass=="L"],
138   col=2,pch=16)
139 points(msdp[carclass=="M"], accelrate[carclass=="M"],
140   col=3,pch=16)

```

```

141 points(msrp[ carclass=="MV" ] , accelrate[ carclass=="MV" ] ,
142   col=4,pch=16)
143 points(msrp[ carclass=="PT" ] , accelrate[ carclass=="PT" ] ,
144   col=5,pch=16)
145 points(msrp[ carclass=="SUV" ] , accelrate[ carclass=="SUV" ] ,
146   col=6,pch=16)
147 points(msrp[ carclass=="TS" ] , accelrate[ carclass=="TS" ] ,
148   col=8,pch=16)
149 abline(a=6.951 , b=1.034e-04 , col=1 , lwd=3) #C
150 abline(a=(6.951+3.224) , b=(1.034e-04 -3.546e-05) ,
151   col=2 , lwd=3) #L
152 abline(a=(6.951+1.818) , b=(1.034e-04 -9.513e-08) ,
153   col=3 , lwd=3) #M
154 abline(a=(6.951-1.144) , b=(1.034e-04 -3.448e-05) ,
155   col=4 , lwd=3) #MV
156 abline(a=(6.951-4.765e-01) , b=(1.034e-04 + 1.988e-05) ,
157   col=5 , lwd=3) #PT
158 abline(a=(6.951+3.082) , b=(1.034e-04 -4.281e-05) ,
159   col=6 , lwd=3) #SUV
160 abline(a=(6.951+6.068) , b=(1.034e-04 -2.665e-04) ,
161   col=8 , lwd=3) #TS
162 legend(20000,35, legend=
163   c("C", "L", "M", "MV", "PT", "SUV", "TS"),
164   col=c(1,2,3,4,5,6,8), pch=c(16,16,16,16,16,16,16),
165   bty = "n", horiz = TRUE)
166
167 ## Final Model msrp + mpg + year:msrp
168 mod.msrmpmg.class.inter = lm(accelrate ~ msrp + mpg +
169                           year:msrp, data=hybrid)
170 par(mfrow=c(2,2))
171 plot(mod.msrmpmg.class.inter)
172 summary(mod.msrmpmg.class.inter)
173 par(mfrow=c(1,1))
174
175 ## Perform a partial F-Test to determine if the interaction term
176 ## makes sense
177 ## Full Final Model
178 Full.model <- lm(accelrate ~ msrp + mpg +
179                      year:msrp, data=hybrid)
180 summary(Full.model)
181 ## Remove the interaction term and make a reduced (nested) model
182 Reduced.model <- lm(accelrate ~ msrp + mpg, data=hybrid)
183 summary(Reduced.model)
184
185 ## Partial F-Test
186 anova(Reduced.model, Full.model)
187
188 ## Two Way Anova

```

```

189 library(car)
190 # Make msrp a categorical variable low < 40k high > 50k
191 hybrid$price <- as.factor(ifelse(hybrid$msrp < 40000, yes="low", no="high"))
192
193
194 summary(hybrid)
195 plot <- ggline(hybrid, x="price", y="accelerate", color="carclass",
196                   add=c("mean_se", "dotplot"))
197 plot
198 model1 <- lm(accelerate~price*carclass, data=hybrid)
199 Anova(model1, type=2)
200
201 ## Check model assumptions
202 ## normal distribution of model residuals
203 plot(model1, 2)
204 aov_residuals <- residuals(object=model1)
205 shapiro.test(aov_residuals)
206
207 ## homogeneity of variance
208 plot(model1, 1)
209 leveneTest(accelerate~price*carclass, data=hybrid)
210
211 ## look at all the separate group combinations
212 posthoc <- lsmeans(model1,
213                       pairwise~price*carclass,
214                       adjust="tukey")
215 posthoc

```