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, accelrate, 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 accelrate 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 `accelrate` 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 `accelrate` and the presence of a good linear correlation between `accelrate` and `msrp`. It also indicated that there might be a linear correlation between `accelrate` and `year` if controlled for `carclass`. In Figure 4, we suspected that there might be a curvilinear relationship between `accelrate` 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 `accelrate`. 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 accelrate 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 shrikage 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 `mgpmpge` 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 `accelrate`.

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 `accelrate`. 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 R^2	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.

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

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 `accelrate` 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 accelrate between the high and low `price` may show an intereaction 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 `accelrate` 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: accelrate
               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
Shapiro-Wilk normality test data: aov_residuals W = 0.97532, p-value = 0.007489	Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F) group 11 1.3613 0.1977 141

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

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

A GRAPHICS AND SUMMARY TABLES

The GLM Procedure

Dependent Variable: accelrate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	108	1251.826921	11.590990	8.20	<.0001
Error	44	62.201033	1.413660		
Corrected Total	152	1314.027954			

R-Square	Coeff Var	Root MSE	accelrate Mean
0.952664	9.942506	1.188974	11.95850

Source	DF	Type I SS	Mean Square	F Value	Pr > F
vehicle	108	1251.826921	11.590990	8.20	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
vehicle	108	1251.826921	11.590990	8.20	<.0001

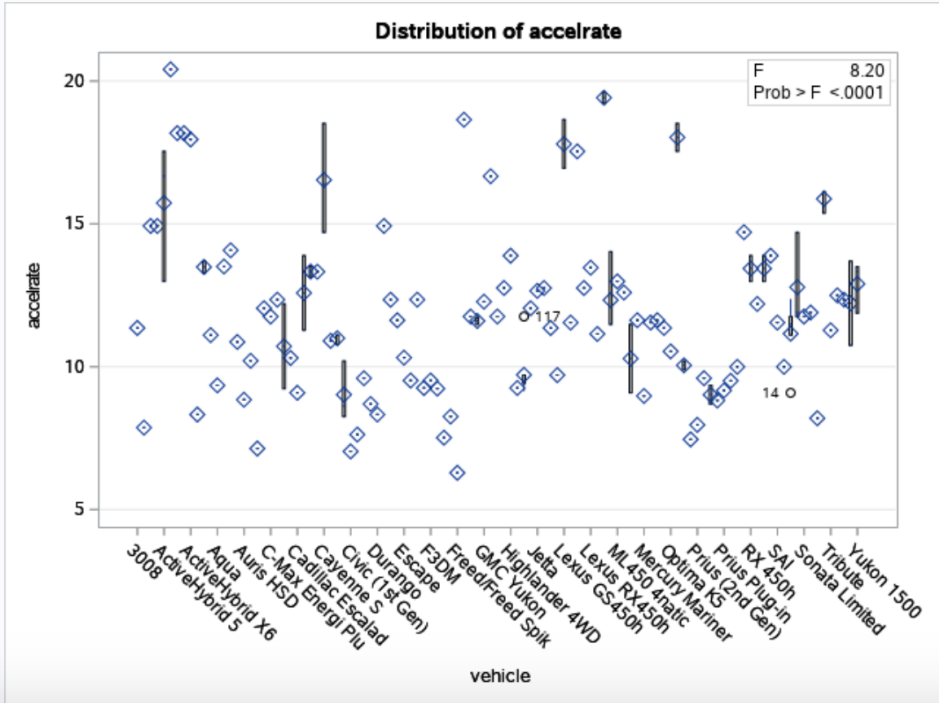


Fig. 1. Accelrate by Vehicle.

The GLM Procedure

Dependent Variable: accelrate

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	accelrate 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

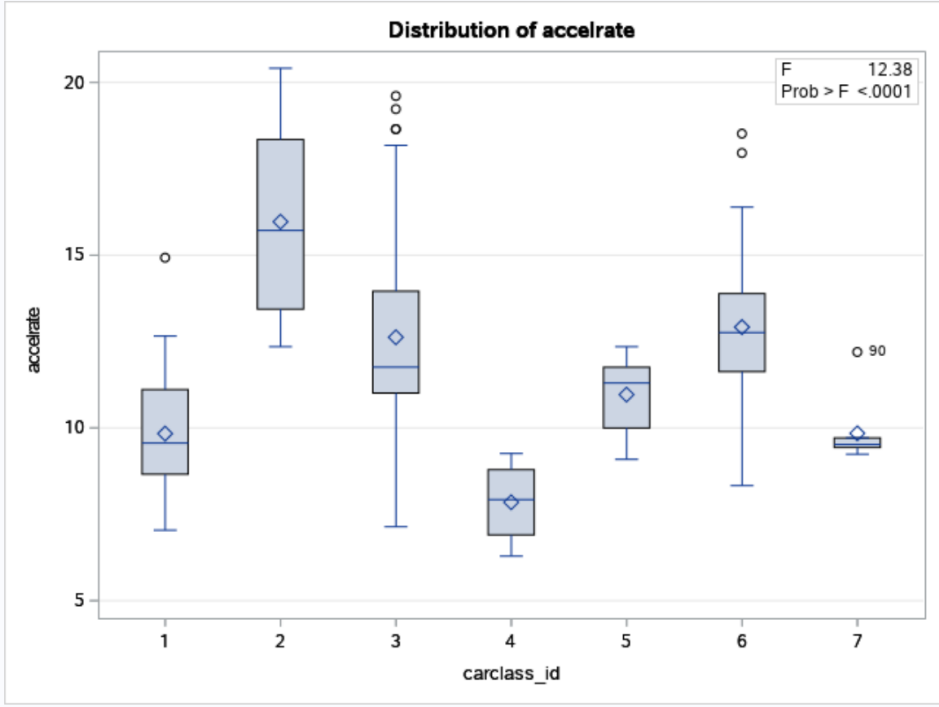


Fig. 2. Accelrate by Car Class.

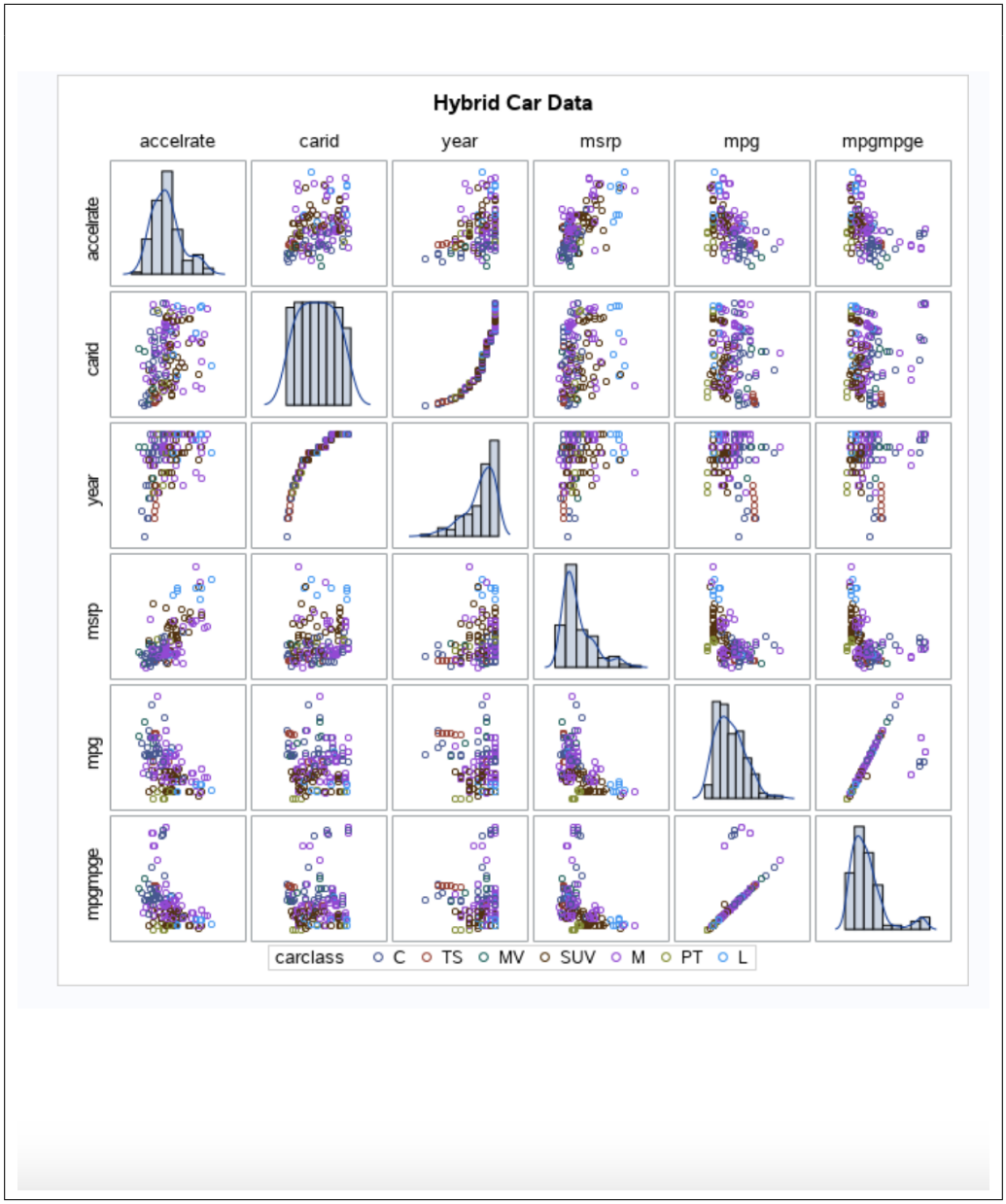


Fig. 3. Scatter Plot Matrix by Car Class.

Simple Statistics

Hybrid Car Data

The CORR Procedure

5 Variables:	accelrate year msrp mpg mpgmpge

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
accelrate	153	11.95850	2.94022	1830	6.29000	20.41000
year	153	2010	3.45927	307487	1997	2013
msrp	153	39319	21421	6015874	11849	118544
mpg	153	34.79745	10.97252	5324	17.00000	72.92000
mpgmpge	153	38.45327	18.36315	5883	17.00000	100.00000

Pearson Correlation Coefficients, N = 153
Prob > |r| under H0: Rho=0

	accelrate	year	msrp	mpg	mpgmpge
accelrate	1.00000	0.35943 <.0001	0.69558 <.0001	-0.50607 <.0001	-0.39887 <.0001
year	0.35943 <.0001	1.00000	0.20979 0.0093	-0.16994 0.0357	0.00549 0.9463
msrp	0.69558 <.0001	0.20979 0.0093	1.00000	-0.53183 <.0001	-0.37222 <.0001
mpg	-0.50607 <.0001	-0.16994 0.0357	-0.53183 <.0001	1.00000	0.66775 <.0001
mpgmpge	-0.39887 <.0001	0.00549 0.9463	-0.37222 <.0001	0.66775 <.0001	1.00000

Scatter Plot Matrix

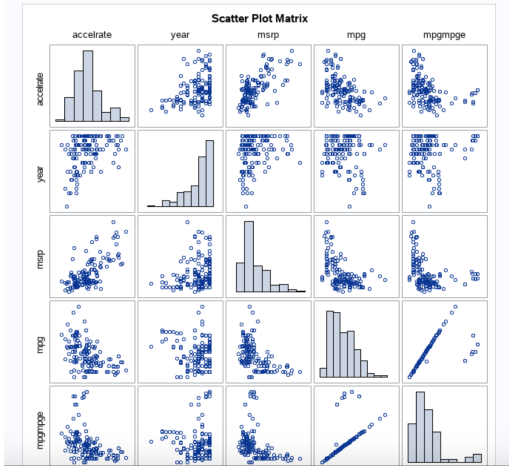


Fig. 4. Hybrid Car Data Correlation.

Simple Statistics

Hybrid Car Data

The CORR Procedure

5 Variables:	accelrate mpg logMPG sqrtMPG MPG2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
accelrate	153	11.95850	2.94022	1830	6.29000	20.41000
mpg	153	34.79745	10.97252	5324	17.00000	72.92000
logMPG	153	3.50086	0.31420	535.63220	2.83321	4.28936
sqrtMPG	153	5.82793	0.91550	891.67350	4.12311	8.53932
MPG2	153	1330	857.96936	203562	289.00000	5317

Pearson Correlation Coefficients, N = 153
Prob > |r| under H0: Rho=0

	accelrate	mpg	logMPG	sqrtMPG	MPG2
accelrate	1.00000	-0.50607 <.0001	-0.50160 <.0001	-0.50729 <.0001	-0.48312 <.0001
mpg	-0.50607 <.0001	1.00000	0.98312 <.0001	0.99567 <.0001	0.98225 <.0001
logMPG	-0.50160 <.0001	0.98312 <.0001	1.00000	0.99584 <.0001	0.93257 <.0001
sqrtMPG	-0.50729 <.0001	0.99567 <.0001	0.99584 <.0001	1.00000	0.96090 <.0001
MPG2	-0.48312 <.0001	0.98225 <.0001	0.93257 <.0001	0.96090 <.0001	1.00000

Scatter Plot Matrix

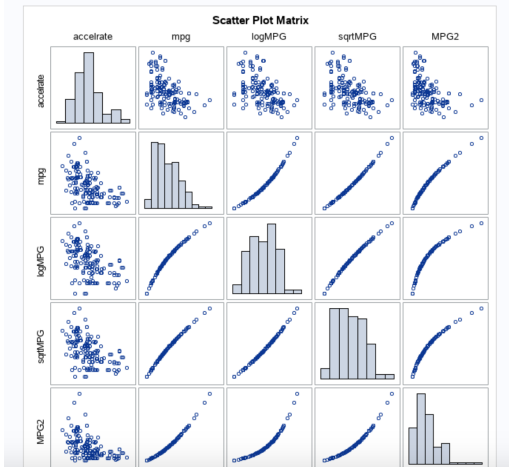


Fig. 5. Transformations of MPG vs accelrate.

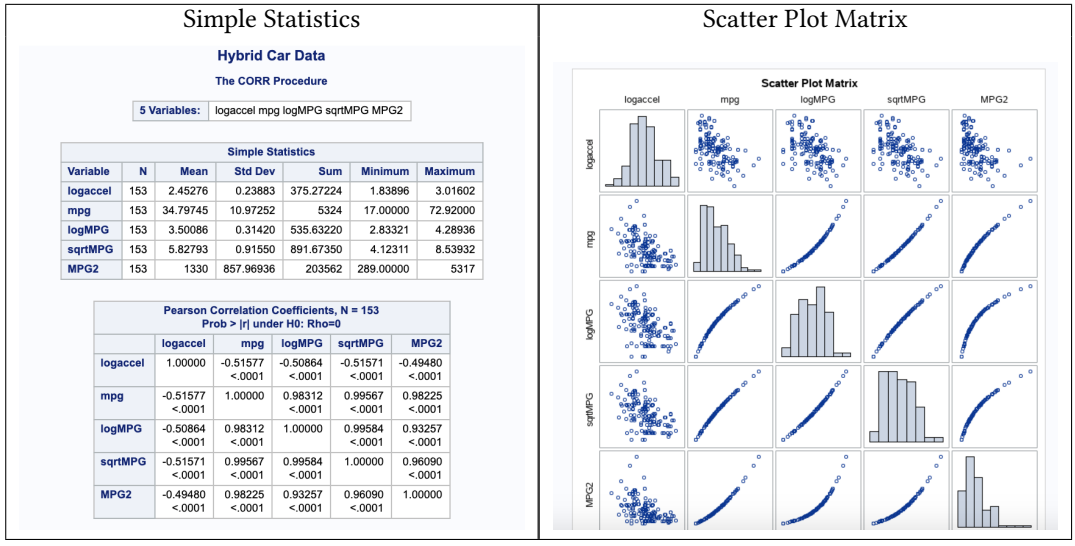


Fig. 6. Transformations of MPG vs logaccelrate.

No Interaction

```
Call:
lm(formula = accelrate ~ mpg + carclass, data = hybrid)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9992 -1.5819 -0.0597  1.1935  6.2141

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.74030    1.17037  12.595  < 2e-16 ***
mpg         -0.11259    0.02521  -4.466 1.59e-05 ***
carclassL    4.16459    1.00963   4.125 6.22e-05 ***
carclassM    1.92070    0.54516   3.523 0.000571 ***
carclassMV  -1.34222    1.22721  -1.094 0.275893
carclassPT  -1.61568    1.19274  -1.355 0.177652
carclassSUV  1.10864    0.70410   1.575 0.117543
carclassTS   0.59464    0.91778   0.648 0.518066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.298 on 145 degrees of freedom
Multiple R-squared:  0.4173,    Adjusted R-squared:  0.3892
F-statistic: 14.84 on 7 and 145 DF,  p-value: 1.663e-14
```

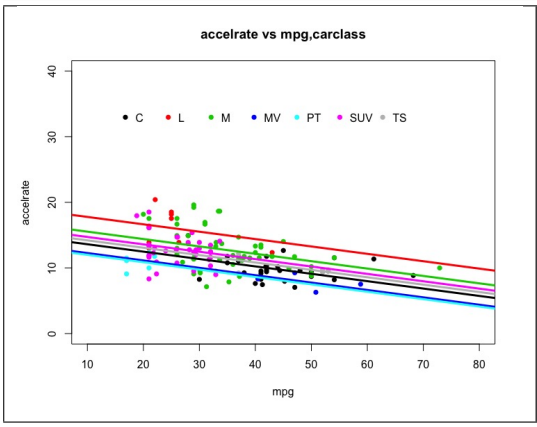
With Interaction

```
Call:
lm(formula = accelrate ~ mpg + carclass + mpg:carclass, data = hybrid)

Residuals:
   Min     1Q Median     3Q    Max
-6.119 -1.375 -0.122  1.252  6.035

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.02325    2.24837   5.348 3.57e-07 ***
mpg            -0.05019    0.05077  -0.988   0.3246
carclassL       8.36455    3.99642   2.093   0.0382 *
carclassM       5.58129    2.58719   2.157   0.0327 *
carclassMV     -0.55122    8.99164  -0.061   0.9512
carclassPT     -4.80128    8.63806  -0.556   0.5792
carclassSUV     4.53937    3.11915   1.455   0.1478
carclassTS      1.42846    6.35540   0.225   0.8225
mpg:carclassL  -0.11931    0.13291  -0.898   0.3709
mpg:carclassM  -0.08874    0.06148  -1.443   0.1511
mpg:carclassMV -0.02332    0.18231  -0.128   0.8984
mpg:carclassPT  0.24556    0.43530   0.564   0.5736
mpg:carclassSUV -0.08984   0.09635  -0.932   0.3527
mpg:carclassTS -0.02377    0.13101  -0.181   0.8563
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.321 on 139 degrees of freedom
Multiple R-squared:  0.4299,    Adjusted R-squared:  0.3766
F-statistic: 8.064 on 13 and 139 DF,  p-value: 7.042e-12
```



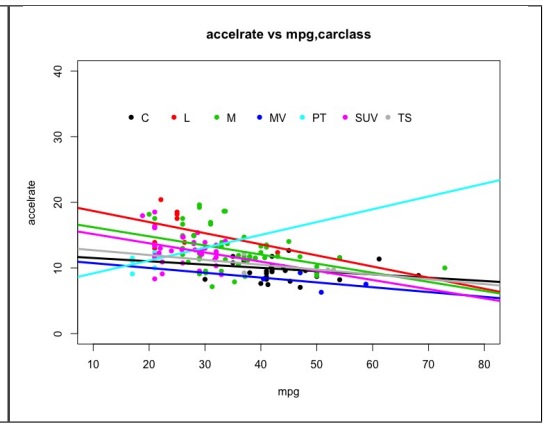


Fig. 7. Interactions between MPG and Car Class on Accelrate.

No Interaction

```
Call:
lm(formula = accelrate ~ msrp + carclass, data = hybrid)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5460 -1.1528 -0.0293  1.1339  5.4715

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.418e+00  4.356e-01  17.030  < 2e-16 ***
msrp         8.672e-05  9.510e-06   9.120 5.76e-16 ***
carclassL    1.160e+00  9.454e-01   1.227  0.22180
carclassM    1.973e+00  4.421e-01   4.461 1.62e-05 ***
carclassMV  -2.136e+00  1.036e+00  -2.061  0.04105 *
carclassPT   3.901e-01  8.730e-01   0.447  0.65562
carclassSUV  1.373e+00  5.023e-01   2.733  0.00706 **
carclassTS   7.404e-01  7.765e-01   0.954  0.34187
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.954 on 145 degrees of freedom
Multiple R-squared:  0.5788,    Adjusted R-squared:  0.5584
F-statistic: 28.46 on 7 and 145 DF,  p-value: < 2.2e-16
```

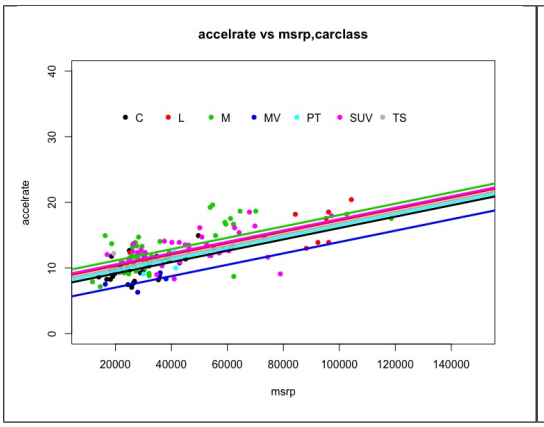
With Interaction

```
Call:
lm(formula = accelrate ~ msrp + carclass + msrp:carclass, data = hybrid)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5075 -1.0707  0.0350  0.9688  5.1819

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.951e+00  1.134e+00   6.132 8.46e-09 ***
msrp              1.034e-04  3.866e-05   2.676  0.00836 **
carclassL         3.224e+00  2.860e+00   1.127  0.26164
carclassM         1.818e+00  1.263e+00   1.440  0.15220
carclassMV       -1.144e+00  3.729e+00  -0.307  0.75952
carclassPT       -4.765e-01  6.954e+00  -0.069  0.94547
carclassSUV       3.082e+00  1.441e+00   2.138  0.03424 *
carclassTS        6.068e+00  1.783e+01   0.340  0.73406
msrp:carclassL   -3.546e-05  4.876e-05  -0.727  0.46833
msrp:carclassM   -9.513e-08  4.084e-05  -0.002  0.99814
msrp:carclassMV  -3.448e-05  1.216e-04  -0.284  0.77717
msrp:carclassPT   1.988e-05  1.910e-04   0.104  0.91725
msrp:carclassSUV -4.281e-05  4.243e-05  -1.009  0.31479
msrp:carclassTS  -2.665e-04  9.141e-04  -0.292  0.77110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.964 on 139 degrees of freedom
Multiple R-squared:  0.5921,    Adjusted R-squared:  0.5539
F-statistic: 15.52 on 13 and 139 DF,  p-value: < 2.2e-16
```



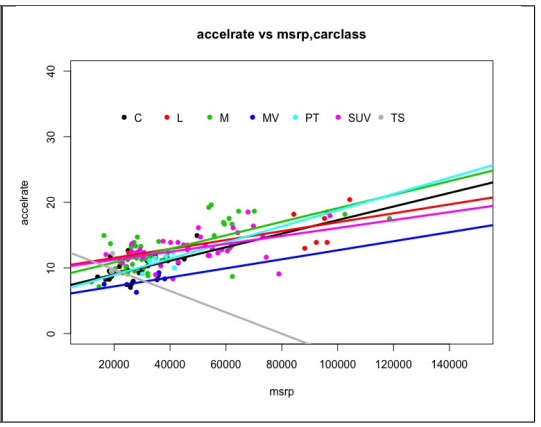


Fig. 8. Interactions between MSPR and Car Class on Accelrate.

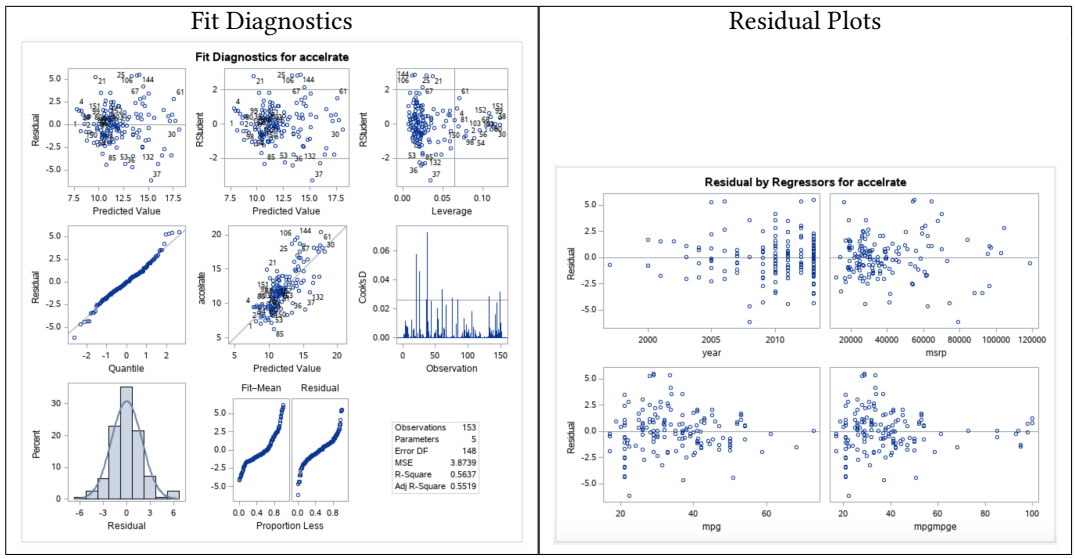


Fig. 9. Assumption Diagnostics for Hybrid Dataset.

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

Model 1: accelrate ~ msrp + mpg
Model 2: accelrate ~ 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.

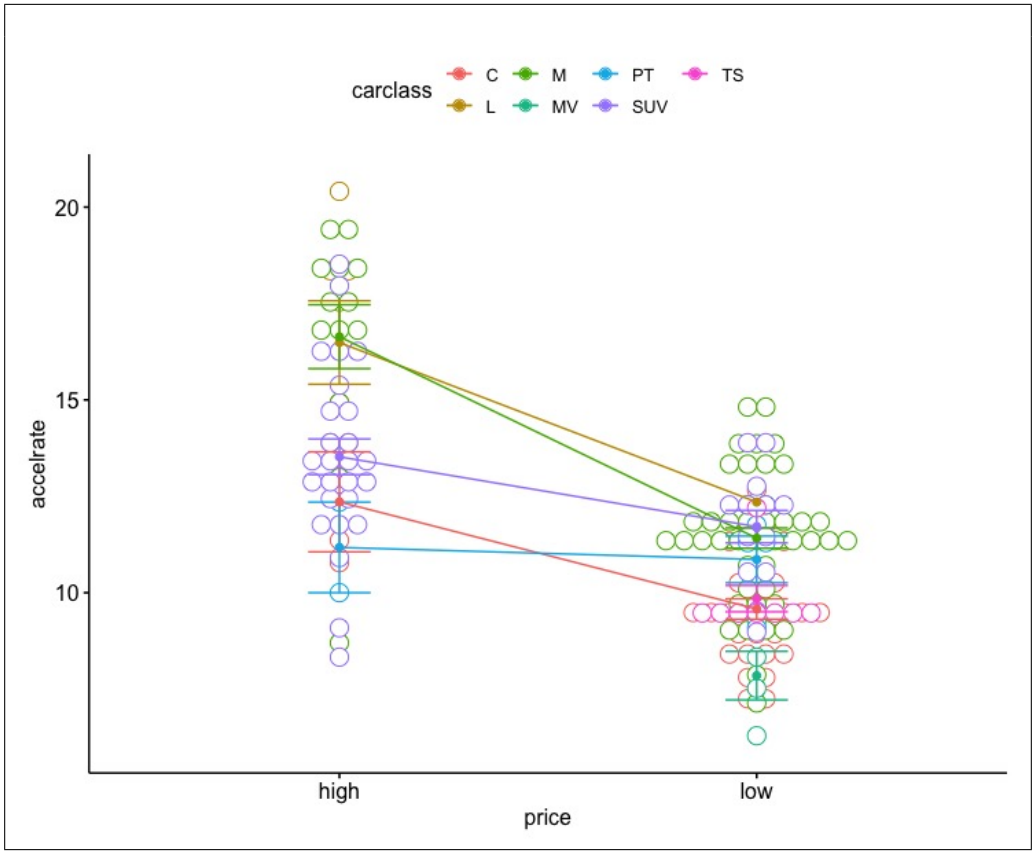


Fig. 11. Plot of Accelrate vs Price by Car Class.

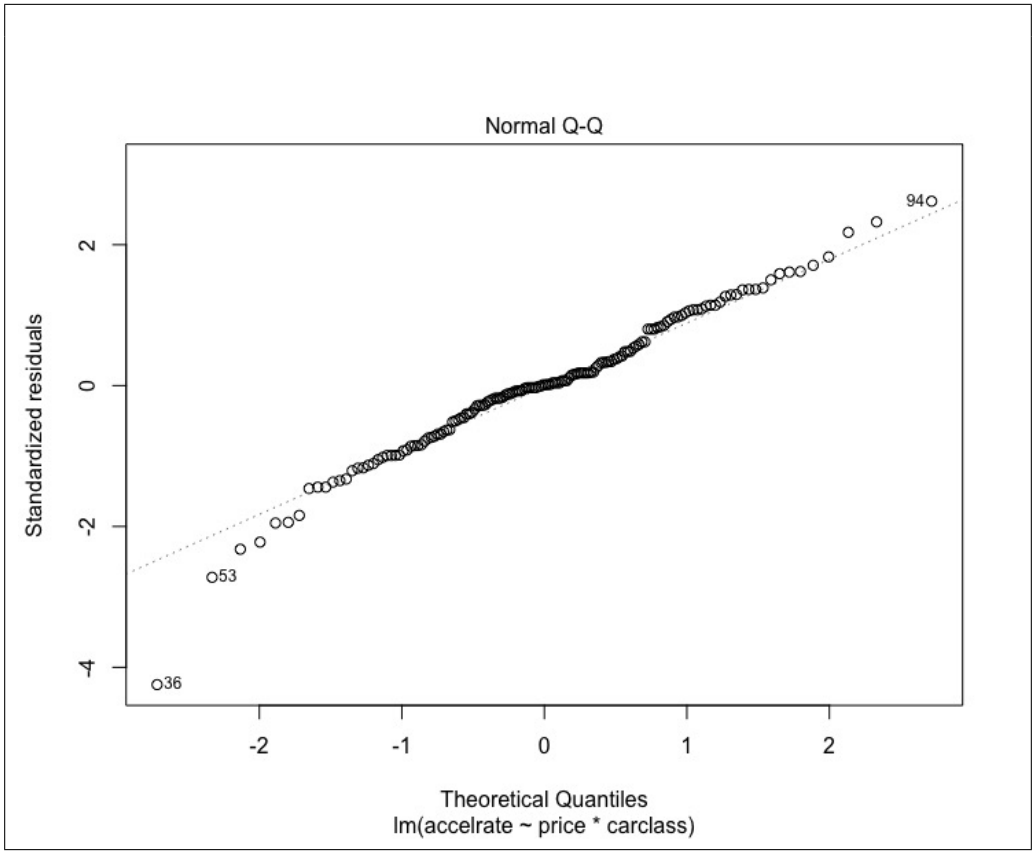


Fig. 12. QQ Plot of Model Residuals.

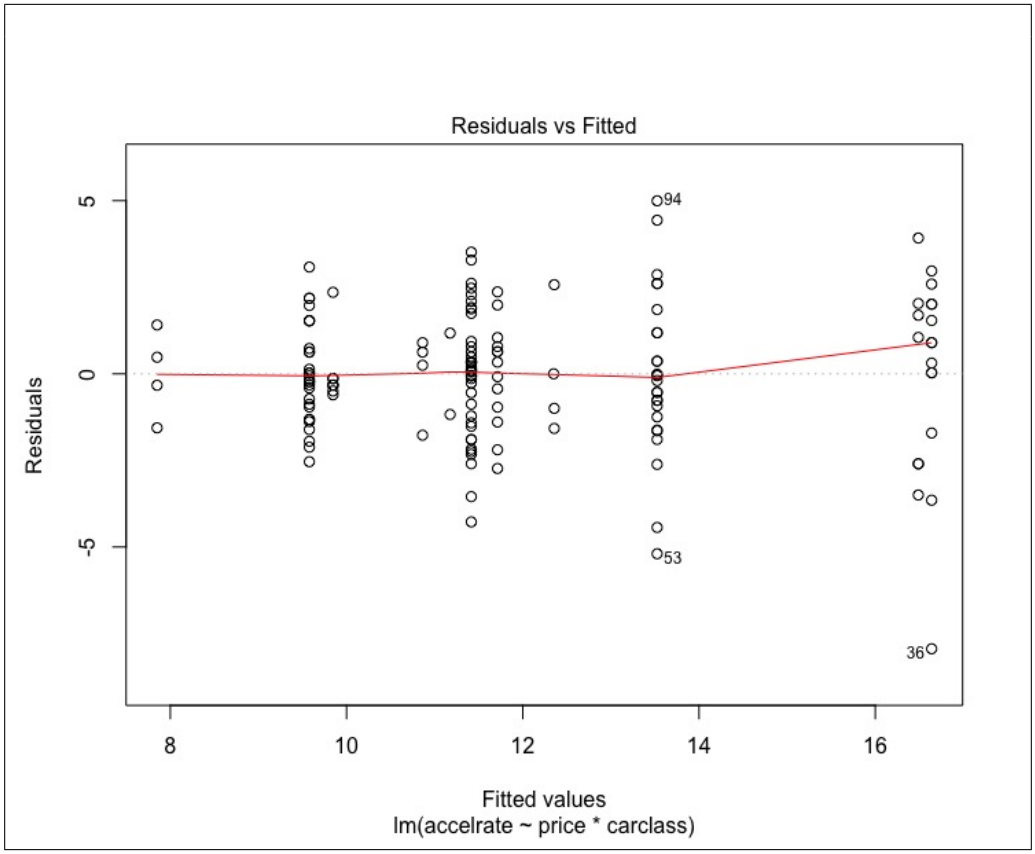


Fig. 13. Fitted Model Residuals.

```
$contrasts
 contrast           estimate    SE  df t.ratio p.value
 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      0.00667 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     4.50667 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      -1.83800 0.468 141  -3.930 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   2.96396 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 2.064 141   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   3.11462 0.661 141   4.711 0.0005
 high,M - low,SUV    4.92615 0.763 141   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).

```
proc import datafile="/home/sarellano0/dataSets/hybrid_reg.csv"
          dbms=dlm out=Hybrid replace;
     delimeter=',';
     getnames=yes;

run;

/* accelrate ~ carid vehicle year msrp mpg mpgmpge carclass carclass_id */

/* bunch of data transformations */
data Hybrid2;
set Hybrid;
logaccel = log(accelrate);
logMPG = log(mpg);
logYear = log(year);
GP100 = 100/mpg;
logGP100 = log(100/mpg);
logMSRP = log(msrp);
sqrtMPG = sqrt(mpg);
sqrtMSRP = sqrt(msrp);
MPG2 = (mpg*mpg);
MSRP2 = (msrp*msrp);
run;


/* EDA of Statisitical Significance of year as Categorical Variable */
proc GLM data=Hybrid2;
        class year;
        model accelrate=year;
        run;

/*EDA of Statisitical Significance of carclass_id as Categorical Variable
proc GLM data=Hybrid2;
        class carclass_id;
        model accelrate=carclass_id;
        run;

/* EDA of Continous Variables */
proc sgscatter data=Hybrid2;
  title "Hybrid Car Data";
  matrix accelrate carid year msrp mpg mpgmpge / group=carclass diagonal=(
run;
  title;

```

```
/* EDA Correlation of Continous Variables*/
ods graphics on;
title 'Hybrid Car Data';
proc corr data=Hybrid2 nomiss plots=matrix(histogram);
   var accelrate year msrp mpg mpgmpge;
 run;
ods graphics off;

/* EDA Correlation of MGP Transformations*/
ods graphics on;
title 'Hybrid Car Data';
proc corr data=Hybrid2 nomiss plots=matrix(histogram);
   var accelrate mpg logmpg sqrtMPG MPG2 ;
 run;
ods graphics off;

/* EDA Correlation of MGP Transformations on logaccel*/
ods graphics on;
title 'Hybrid Car Data';
proc corr data=Hybrid2 nomiss plots=matrix(histogram);
   var logaccel mpg logmpg sqrtMPG MPG2 ;
 run;
ods graphics off;

/* EDA examining data distribution of mpg */
proc univariate data = Hybrid2;
var mpg;
histogram;
run;

/* Linear Modeling Assumptions Continous Variables with VIF */
proc reg data = Hybrid2 corr plots(label) = all;
        model accelrate = year msrp mpg mpgmpge/ VIF /*CLB*/;
        title 'Hybrid Car Data';
        run; quit;

/* Linear Modeling Assumptions All Variables */
proc glm data = Hybrid2 plots=all;
  class carclass_id;
  model accelrate = year msrp mpg mpgmpge carclass_id;
run;
quit;

/* Model Selection simple glmselect with no interaction variables Stepwise
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id / sele
        run; quit; /* year msrp mpgmpge
```

```
                                RMSE 1.968 Adj R-Sq     0.552 AICC 366.474

/* Model Selection simple glmselect with no interaction variables LARS no
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id / sele
        run; quit; /* msrp
                                RMSE    2.286 Adj R-Sq  0.395 AICC 410.205

/* Model Selection simple glmselect with no interaction variables LASSO no
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = year msrp mpg mpgmpge carclass_id / selection =L
        run; quit; /* msrp
                                RMSE    2.286 Adj R-Sq  0.395 AICC 410.205

/* Model Selection simple glmselect with no interaction variables ELASTICN
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = year msrp mpg mpgmpge carclass_id / selection =E
        run; quit; /* year msrp mpg mpgmpge carclass_id_1 carclass_id_2 ca
                                RMSE 1.961 Adj R-Sq     .550 AICC 372.661


/* Model Selection simple glmselect with Interaction variables Stepwise no
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection =STEPWISE;
        run; quit; /* msrp mpgmpge year*msrp
                                RMSE 1.950 Adj R-Sq 0.569 AICC 363.652 SBC

/* Model Selection simple glmselect with Interaction variables LAR no CV*/
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection =LAR;
        run; quit; /* year*msrp
                                RMSE 2.284 Adj R-Sq     0.397 AICC 409.673

/* Model Selection simple glmselect with no interaction variables LASSO no
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection =LASSO;
        run; quit; /* year*msrp
                                RMSE 2.284 Adj R-Sq     0.397 AICC 409.673

```

```
/* Model Selection simple glmselect with no interaction variables ELEASTIC
proc glmselect data = Hybrid2;
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection =ELASTICNET;
        run; quit; /* year*msrp
                                RMSE 2.284 Adj R-Sq     0.397 AICC 409.673


/* Model Selection glmselect with Interaction variables Stepwise with CV*/
proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A
        partition fraction(test = .25 validate = .25);
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection = stepwise( stop = CV) cvdetails;
        run; quit; /* msrp mpgmpge year*msrp
                                RMSE 1.819 Adj R-Sq 0.619 AICC 182.429 SBC

/* Model Selection glmselect with Interaction variables LAR with CV*/
proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A
        partition fraction(test = .25 validate = .25);
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection = LAR( stop = CV) cvdetails;
        run; quit; /* mpg year*msrp
                                RMSE 2.041 Adj R-Sq 0.520 AICC 199.653 SBC

/* Model Selection glmselect with Interaction variables LASSO with CV*/
proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A
        partition fraction(test = .25 validate = .25);
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection = LASSO( stop = CV) cvdetails;
        run; quit; /* mpg year*msrp
                                RMSE 2.041 Adj R-Sq 0.520 AICC 199.653 SBC

/* Model Selection glmselect with Interaction variables ELASTICNET with CV
proc glmselect data = Hybrid2 plots(stepaxis = number) = (criterionpanel A
        partition fraction(test = .25 validate = .25);
        class carclass_id vehicle;
        model accelrate = vehicle year msrp mpg mpgmpge carclass_id msrp*y
        selection = ELASTICNET( stop = CV) cvdetails;
        run; quit; /* msrp mpg year*msrp
                                RMSE 2.055 Adj R-Sq 0.513 AICC 201.986 SBC

/* Final Model Fit Diagnostics */
proc glm data = Hybrid2 plots=all;
  model accelrate = msrp mpg year*msrp;
```

```
189 run;
190 quit;
```

C R CODE

Listing 2. R Code (AppliedStatsProject.R).

```
#Load dataset
hybrid<-read.csv("hybrid_reg.csv")
summary(hybrid)
## accelrate with mpg and carclass - no interaction
mod.mpg.class.nointer = lm(accelrate ~ mpg + carclass,data=hybrid)
par(mfrow=c(2,2))
plot(mod.mpg.class.nointer)
summary(mod.mpg.class.nointer)
par(mfrow=c(1,1))
plot(mpg[carclass=="C"], accelrate[carclass=="C"], col=1, pch=16,
     xlab="mpg", ylab="accelrate",
     main="accelrate vs mpg,carclass",
     ylim=c(0,40), xlim=c(10,80))
points(mpg[carclass=="L"], accelrate[carclass=="L"],
       col=2,pch=16)
points(mpg[carclass=="M"], accelrate[carclass=="M"],
       col=3,pch=16)
points(mpg[carclass=="MV"], accelrate[carclass=="MV"],
       col=4,pch=16)
points(mpg[carclass=="PT"], accelrate[carclass=="PT"],
       col=5,pch=16)
points(mpg[carclass=="SUV"], accelrate[carclass=="SUV"],
       col=6,pch=16)
points(mpg[carclass=="TS"], accelrate[carclass=="TS"],
       col=8,pch=16)
abline(a=14.74030, b=-0.11259, col=1, lwd=3) #C
abline(a=(14.74030+4.16459), b=-0.11259,
       col=2, lwd=3) #L
abline(a=(14.74030+1.92070), b=-0.11259,
       col=3, lwd=3) #M
abline(a=(14.74030-1.34222), b=-0.11259,
       col=4, lwd=3) #MV
abline(a=(14.74030-1.61568), b=-0.11259,
       col=5, lwd=3) #PT
abline(a=(14.74030+1.10864), b=-0.11259,
       col=6, lwd=3) #SUV
abline(a=(14.74030+0.59464), b=-0.11259,
       col=8, lwd=3) #TS
legend(15,35, legend=
         c("C", "L", "M", "MV", "PT", "SUV", "TS"),
       col=c(1,2,3,4,5,6,8),
       pch=c(16,16,16,16,16,16,16),
       bty = "n",horiz=TRUE)

```

```
## accelrate with mpg and carclass - interaction
mod.mpg.class.inter = lm(accelrate ~ mpg + carclass +
                           mpg:carclass,
                         data=hybrid)
par(mfrow=c(2,2))
plot(mod.mpg.class.inter)
summary(mod.mpg.class.inter)
par(mfrow=c(1,1))
plot(mpg[carclass=="C"], accelrate[carclass=="C"],
     col=1, pch=16,
     xlab="mpg", ylab="accelrate",
     main="accelrate vs mpg,carclass",
     ylim=c(0,40), xlim=c(10,80))
points(mpg[carclass=="L"], accelrate[carclass=="L"],
       col=2,pch=16)
points(mpg[carclass=="M"], accelrate[carclass=="M"],
       col=3,pch=16)
points(mpg[carclass=="MV"], accelrate[carclass=="MV"],
       col=4,pch=16)
points(mpg[carclass=="PT"], accelrate[carclass=="PT"],
       col=5,pch=16)
points(mpg[carclass=="SUV"], accelrate[carclass=="SUV"],
       col=6,pch=16)
points(mpg[carclass=="TS"], accelrate[carclass=="TS"],
       col=8,pch=16)
abline(a=12.02325, b=-0.05019, col=1, lwd=3) #C
abline(a=(12.023258 + 8.36455), b=(-0.05019 -0.11931),
       col=2, lwd=3) #L
abline(a=(12.02325+5.58129), b=(-0.05019 -0.08874),
       col=3, lwd=3) #M
abline(a=(12.02325-0.55122), b=(-0.05019 -0.02332),
       col=4, lwd=3) #MV
abline(a=(12.02325-4.80128), b=(-0.05019 + 0.24556),
       col=5, lwd=3) #PT
abline(a=(12.02325+4.53937), b=(-0.05019 -0.08984),
       col=6, lwd=3) #SUV
abline(a=(12.02325+1.42846), b=(-0.05019 -0.02377),
       col=8, lwd=3) #TS
legend(15,35, legend=
         c("C", "L", "M", "MV", "PT", "SUV", "TS"),
       col=c(1,2,3,4,5,6,8), pch=c(16,16,16,16,16,16,16),
       bty = "n",horiz=TRUE)

## accelrate with msrp and carclass - no interaction
mod.msrp.class.nointer = lm(accelrate ~ msrp + carclass,
                            data=hybrid)
par(mfrow=c(2,2))
plot(mod.msrp.class.nointer)
```

```
summary(mod.msrp.class.nointer)
par(mfrow=c(1,1))
plot(msrp[carclass=="C"], accelrate[carclass=="C"],
     col=1, pch=16,
     xlab="msrp", ylab="accelrate",
     main="accelrate vs msrp,carclass",
     ylim=c(0,40), xlim=c(10000,150000))
points(msrp[carclass=="L"], accelrate[carclass=="L"],
       col=2,pch=16)
points(msrp[carclass=="M"], accelrate[carclass=="M"],
       col=3,pch=16)
points(msrp[carclass=="MV"], accelrate[carclass=="MV"],
       col=4,pch=16)
points(msrp[carclass=="PT"], accelrate[carclass=="PT"],
       col=5,pch=16)
points(msrp[carclass=="SUV"], accelrate[carclass=="SUV"],
       col=6,pch=16)
points(msrp[carclass=="TS"], accelrate[carclass=="TS"],
       col=8,pch=16)
abline(a=7.418, b=8.672e-05, col=1, lwd=3) #C
abline(a=(7.418+1.160), b=8.672e-05, col=2, lwd=3) #L
abline(a=(7.418+1.973), b=8.672e-05, col=3, lwd=3) #M
abline(a=(7.418-2.136), b=8.672e-05, col=4, lwd=3) #MV
abline(a=(7.418+3.901e-01), b=8.672e-05, col=5, lwd=3) #PT
abline(a=(7.418+1.373), b=8.672e-05, col=6, lwd=3) #SUV
abline(a=(7.418+7.404e-01), b=8.672e-05, col=8, lwd=3) #TS
legend(20000,35, legend=
         c("C", "L", "M", "MV", "PT", "SUV", "TS"),
       col=c(1,2,3,4,5,6,8), pch=c(16,16,16,16,16,16,16),
       bty = "n", horiz=TRUE)

## accelrate with msrp and carclass - interaction
mod.msrp.class.inter = lm(accelrate ~ msrp + carclass +
                            msrp:carclass,
                          data=hybrid)
par(mfrow=c(2,2))
plot(mod.msrp.class.inter)
summary(mod.msrp.class.inter)
par(mfrow=c(1,1))
plot(msrp[carclass=="C"], accelrate[carclass=="C"],
     col=1, pch=16,
     xlab="msrp", ylab="accelrate",
     main="accelrate vs msrp,carclass",
     ylim=c(0,40), xlim=c(10000,150000))
points(msrp[carclass=="L"], accelrate[carclass=="L"],
       col=2,pch=16)
points(msrp[carclass=="M"], accelrate[carclass=="M"],
       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.msrpmpg.class.inter = lm(accelrate ~ msrp + mpg +
169                                 year:msrp, data=hybrid)
170 par(mfrow=c(2,2))
171 plot(mod.msrpmpg.class.inter)
172 summary(mod.msrpmpg.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
```

```
library(car)
# Make msrp a categorical variable low < 40k high > 50k
hybrid$price <- as.factor(ifelse(hybrid$msrp < 40000, yes="low", no="high"
)


summary(hybrid)
plot <- ggline(hybrid, x="price", y="accelrate", color="carclass",
               add=c("mean_se", "dotplot"))
plot
model1 <-lm(accelrate~price*carclass, data=hybrid)
Anova(model1,type=2)

## Check model assumptions
## normal distribution of model residuals
plot(model1, 2)
aov_residuals <- residuals(object=model1)
shapiro.test(aov_residuals)

## homogeneity of variance
plot(model1, 1)
leveneTest(accelrate~price*carclass, data=hybrid)

## ook at all the separate group combinations
posthoc <- lsmeans(model1,
                   pairwise~price*carclass,
                   adjust="tukey")
posthoc
```