PSTAT 126 Project

Hybrids

Ryan Bernstein

Sophie Campione

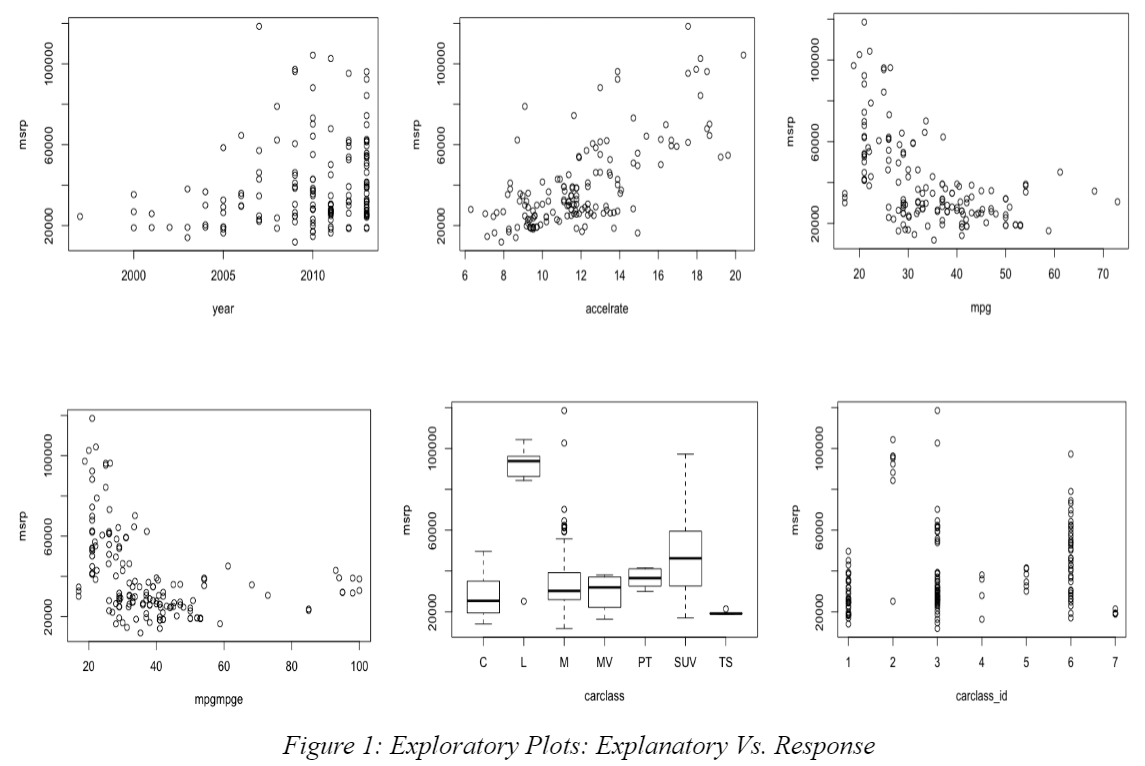
Ann Shi

Nuan Wen

**Abstract**

We were tasked with constructing a linear regression model that accurately predicts the MSRP of a hybrid car. Our dataset comprises of 153 out of 154 unique models (with observation 136 missing), each with five explanatory variables that we were permitted to utilize. The five variables are: model year, acceleration rate (in km/hour/second), fuel economy (in MPG), alternative fuel economy for electric cars (in MPGe), and carclass, which is a categorical abbreviation that separates our models into seven unique classes: Compact, Midsize, 2 Seater, Large, Pickup Truck, Minivan and SUV. We assume that the MSRP values, or response variables, are normally distributed and independent. Our final model has explanatory variables of Accelrate, MPG and regrouped Carclass. Variable selection implies that Year and MPGe are insignificant to our response variable, MSRP.

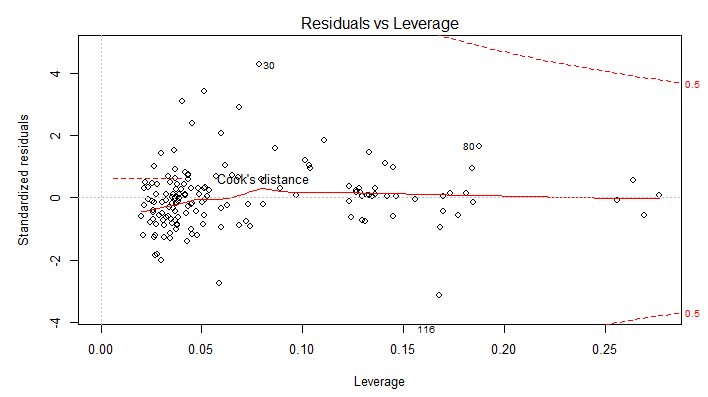
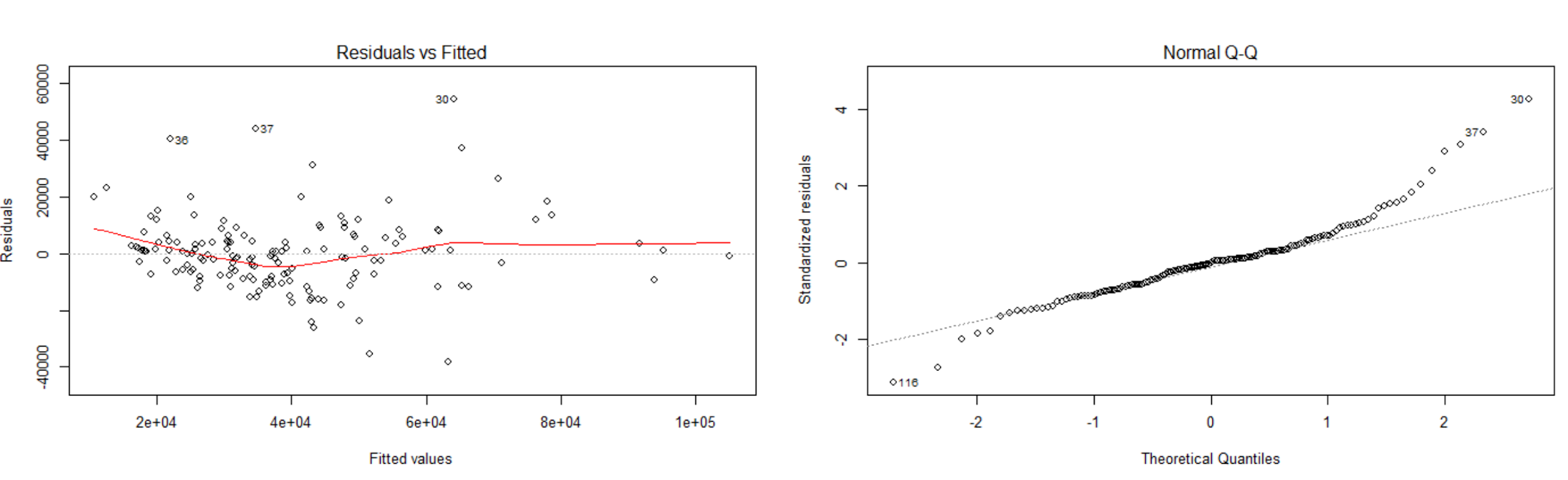
**Methods**

Preliminary analysis helps determine the relationships between the predictor variables and the response variable. *Figure 1* contains the plots of the explanatory variables versus the response variable (MSRP) of the hybrid cars. The upper-left and middle graphs show a widening positive linear relationship between Year and MSRP, and the one between acceleration rate and MSRP, with a clustering towards slower acceleration and cheaper cars. The top-right and bottom-left graphs show a curved, but sloping negative relationship between MPG and MSRP and a similar relationship between MSRP and the maximum of MPGe. The plot in the lower-middle shows the Carclass versus the MSRP. We can see that large cars are the most expensive, since they are clustered with the highest MSRP. SUVs and mid-sized cars have the largest spread. Two-seaters all cost approximately the same, as they have the smallest spread. Formal tests are also used to analyze the explanatory variables.

The initial model that we selected consists of all the explanatory variables, such that the MSRP of the hybrid vehicle is a linear combination of variables that represent vehicle characteristics, plus the random error. The random error term follows a normal distribution with mean 0 and is assumed to have constant variance.

Initial model: MSRPi = β0 + β1 Yeari + β2 Accelratei + β3 MPGi + β4 MPGMPGei + β5 Carclassi + ɛi,

where β0 is the intercept, β1, β2, β3, β4, and β5 are the coefficients for the effect of the predictor variables on the response variable (MSRP), and ɛi, is the random error term (i=1,…,153).

**

*Figure 2*

Diagnostics were then performed to show how well the initial model (and following models) fit the data. The residuals vs. fitted and scale-location plots check the assumptions of constant variance and constant error, which is assumed to be distributed evenly around 0. Looking at the left graph in *Figure 2*, we did not see a random pattern and concluded that our assumptions were invalid. From the normal Q-Q plot, we see that the line does not fit the data well, which violates our assumption of linearity. The residuals vs. leverage plot helps determine influential data points. Influential points can alter the coefficients, and thus the fit of the model, if they do not follow the pattern of the other data points. Therefore, they are often removed to improve fit. From the right plot, we do not observe any points of particular concern, but we need to consider the influence of large leverage points on future models.

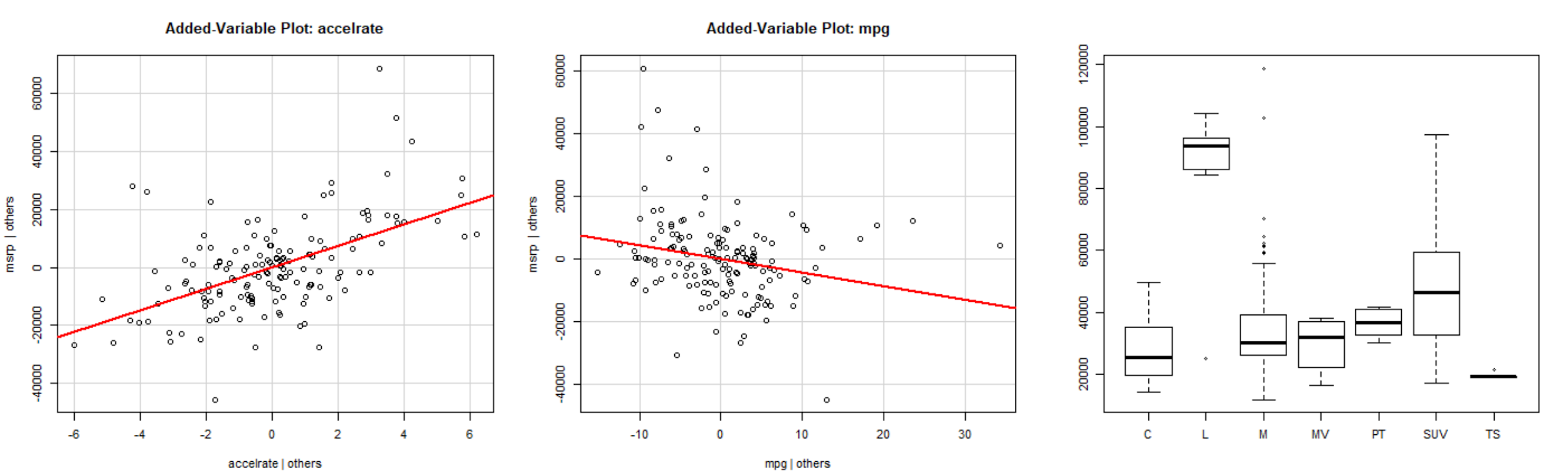
Diagnostic plots show the initial model was a bad fit. An improved model would have a closer fit, and a lower amount of complexity. To figure out how to improve the model, we first analyzed the pairwise plots to observe if any collinearity existed between predictor variables. Interaction terms will greatly influence our interpretation. Variables that appear insignificant can be excluded from the model each at a time. But if interaction terms or high order terms are added, the previously excluded variables are included because they may become significant.

Added variable plots and marginal model plots were used to see if the model was a good fit. Significant p-values and F-statistics were used to reduce the model. We used AIC and BIC to verify that our reduced model is the best model we can get with no transformation, and to ensure the model does not under fit, nor over fit the data. AIC balances goodness of fit with complexity by penalizing added predictor variables and favoring a close fit. BIC favors simpler models even more heavily than AIC. If the new model did not fit well in the diagnostics, we would go on to perform a transformation as suggested by the Box-Cox method to the explanatory variables or the response variable. This would add complexity to the analysis, but transformations can fix violated assumptions such as normality, constant variance and linearity. Diagnostics will be performed on each proposed model, in order to assure that the assumptions remain valid under the different models, and to identify influential points.

**Statistical Analysis**

The statistical analysis was performed via the application R. During our preliminary analysis, see *Figure 1*, the possible underlying relationship between each explanatory variable and the response was plotted. We observed that there might be a correlation between Accelrate and MPG. The explanatory variables “year” and “electric miles per gallon equivalent” (MPGe) appeared to be insignificant. We removed them one at a time to avoid having a significant change in coefficients after removing one variable. The AIC and BIC verified that the model without these two variables was the optimized model. We shifted our analysis to this model, which contains two explanatory variables (accelerate and MPG) and 1 dummy variable (Carclass). We named it Model 2.

Model 2: MSRPi = β0 + β1 Accelratei + β2 MPGi + β3 Carclassi + ɛi, i=1,…,153

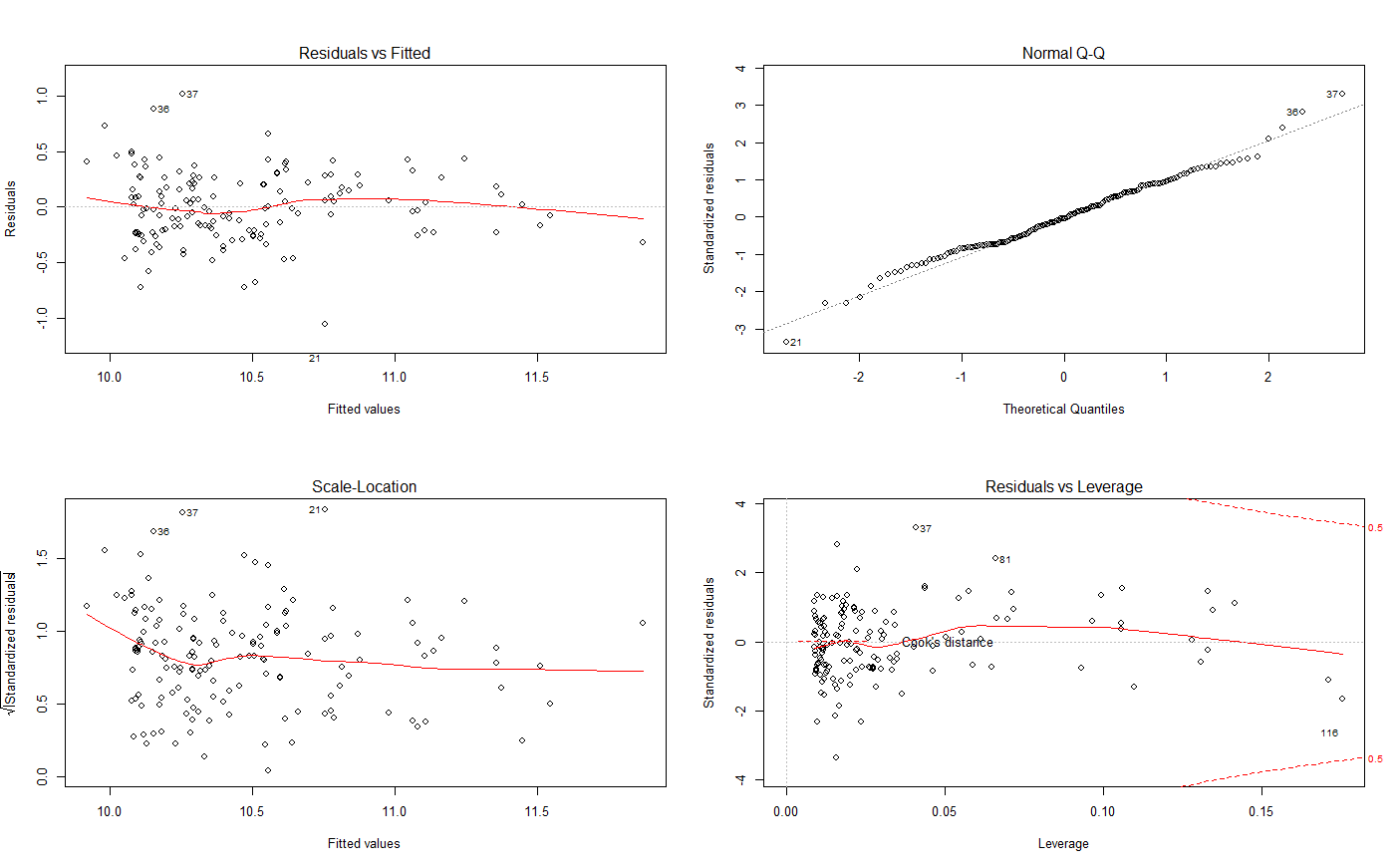


*Figure 3*

By examining the boxplot and added-variable plots in *Figure 3*, we noticed that accelrate and MPG have a strong correlation with MSRP, and the Carclass “large cars” (L) is distinct from other car types. The R summary supported our conclusion. Therefore, we regrouped the Carclass and marked L as 1 and other classes as 0. Then we reexamined the the model and found that all explanatory variables and levels in the categorical variable had become significant. The added-variable plots verify the relationship between the response variable and each of predictor variables (MPG, Accelrate) separately, adjusting for the effects of other x variables. The residuals vs. fitted plot is also improved, but still not randomly distributed.

Next, we examined the model to see if there was a potential interaction between the explanatory variables. Even though VIF does not signal any collinearity, the correlation test and plotting between MPG and Accelrate shows there is an interaction exists. F-test and diagnosis points imply that the model including an interaction term is better. Log transformations were then applied as suggested by Box-Cox. We ran the AIC for a model with a transformed response variable, and ran it again for a different model with transformed response and explanatory variables. We discovered that their AIC and BIC values are close. For simplicity, we chose to only transform the response variable (MSRP) and marked it as Model 3, our final model.

Final model (Model 3): log(MSRPi)= β0 + β1 Accelratei + β2 MPGi + β3 Carclassi + β4 Accelratei\*MPGi +ɛi, where i=1,…,153, Carclass: L = “1”, Others = “0”



*Figure 4*

β0 was the intercept and β1, β2, β3, β4were the effect coefficients. ɛi was assumed to be the same as in the initial model. *Figure 4* has the four diagnostic plots for our model with the transformed response variable. The top-left and bottom-left plots show that our assumptions of constant variance were valid, though variance is slightly downward trending at the beginning. The top-right graph indicates that our observations were in fact normally distributed. The bottom-right graph indicates that observation 116 is high in leverage but is not an influential point. We attempted to verify observation 116 by removing it first, and repeating our previous method of model derivation. The model we arrived at was the same as our final model. We concluded that observation 116 did not significantly affect the prediction coefficient.

**Conclusion**

The explanatory variables that best explain the MSRP for hybrid cars (in US-dollars) are acceleration rate (in km\*h-1/sec), fuel efficiency, (miles travelled burning one gallon of gas, MPG), vehicle type (whether the hybrid is a large car or not) and an interaction term between acceleration rate and fuel efficiency.

The dummy variable Carclass separates our interpretation for the final model into two categories: one for all large hybrid cars (where an additive effect on price is found), and another for all other car types. If all other terms in the final model are hold constant, the price of large hybrid cars will be 34.7% higher than the price of non-large hybrid cars. This result is reasonable, since cars with more room are usually more comfortable for driving and riding, and thus have a higher price.

The rest of the effect coefficients do not have a significant interaction with the dummy variable, so they apply to all types of hybrid cars. Given that acceleration rate (abbr. as AR below) and fuel efficiency (abbr. as FE below) are interacted in our model, we interpret the effect coefficients as the following: For each km\*h-1/sec increase in AR when FE equals 0, the expected price will rise by 19.2%. For each mile/gal increase in FE when AR equals 0, the expected price will rise by 2.7%. Also, for each unit of increase in term AR\*FE, the expected price will drop by 0.3% besides the effects of AR and FE separately. Admittedly, the interaction term included might impair our interpretation of the model. The resulted effect of each term in the model (except Carclass) might be more complex than what we stated above, but we still include it because the overall model gives a more accurate prediction.

These statistics are all consistent with our common knowledge of cars. Faster cars are often more expensive. Meanwhile, many customers are willing to pay more in exchange for higher fuel efficiency (overall MPG). However, higher acceleration generally leads to less fuel economy and customers have to make a choice between high acceleration rate and high fuel efficiency, since they are a pair of zero-sum properties[[1]](#footnote-0), as illustrated by the interaction term. The model shows the correlation between variables, and helps manufacturers find a more balanced combination of acceleration rate and fuel efficiency, allowing them to derive an appropriate retail price for hybrid cars based on different combinations.

**Appendix**

## Preparation

library(car)

library(MASS)

library(alr3)

library(stats4)

library(leaps)

cars = hybrid

cars <- data.frame(hybrid[, c(4,3,5,6,7,8)])

attach(cars)

x.5 <-cbind(year,accelrate, MPG, MPGMPGe, carclass)

cor(x.5)

## Reduce Model (Auto)

FullModel <- lm(msrp~., data = cars)

n = length(FullModel$residuals)

# AIC

stepAIC(FullModel, direction = c("both", "backward", "forward"),

k = 2)

step(FullModel,direction="both", criterion = "BIC", k = log(n))

both.reduced <- lm(formula = msrp ~ accelrate + MPG + carclass, data = cars)

#Forward selection

forwardAIC <- step(FullModel,direction="forward", data=cars)

forwardBIC <- step(FullModel,direction="forward", data=cars,k=log(n))

forward.full <- lm(formula = msrp ~ year + accelrate + MPG + MPGMPGe + carclass,

data = cars)

anova(both.reduced, forward.full) #the result is insignificant, reduced is better (3 variables)

## Adjustion of the Model VIF, PAIR, AV, VIF, MMP, SUMMARY)

# carclass as a whole with identity function

fit.i0 <- lm(msrp ~ year + accelrate + MPG + MPGMPGe + I(carclass))

summary(fit.i0)

pairs(msrp ~ year + accelrate + MPG + MPGMPGe + carclass)

fit = fit.i0

par(mfrow=c(2,2))

plot(fit)

avPlots(fit)

vif(fit) > 5

par(mfrow=c(2,3))

mmp(fit,year)

mmp(fit,accelrate)

mmp(fit,MPG)

mmp(fit,MPGMPGe)

mmp(fit,fit$fitted.values,xlab="Fitted Values")

boxplot(msrp~carclass)

#-MPGe

fit.i-MPGe <- lm(msrp ~ year + accelrate + MPG + I(carclass))

summary(fit.i1)

# reduced model

fit.i1 <- lm(msrp ~ accelrate + MPG + I(carclass))

summary(fit.i1)

par(mfrow = c(2,2))

plot(fit.i1)

avPlots(fit.i1)

par(mfrow = c(2,3))

avPlot(fit.i1,accelrate)

avPlot(fit.i1,MPG)

boxplot(msrp ~ carclass)

par(mfrow = c(2,2))

StanRes1 <- rstandard(fit.i1)

plot(accelrate,StanRes1,ylab="Standardized Residuals")

plot(MPG,StanRes1,ylab="Standardized Residuals")

plot(carclass,StanRes1,ylab="Standardized Residuals")

plot(fit.i1$fitted.values,StanRes1,ylab="Standardized Residuals")

# accelrate vs. msrp (points)

par(mfrow=c(1,1))

plot(accelrate[carclass=="L"],msrp[carclass=="L"],pch=c("L"),col=c("red"),ylab="msrp",xlab="accelrate", xlim = c(min(cars$accelrate), max(cars$accelrate)), ylim = c(min(cars$msrp), max(cars$msrp)))

points(accelrate[carclass=="C"],msrp[carclass=="C"],pch=c("C"),col=c("orange"))

points(accelrate[carclass=="M"],msrp[carclass=="M"],pch=c("M"),col=c("yellow"))

points(accelrate[carclass=="MV"],msrp[carclass=="MV"],pch=c("V"),col=c("green"))

points(accelrate[carclass=="PT"],msrp[carclass=="PT"],pch=c("P"),col=c("blue"))

points(accelrate[carclass=="SUV"],msrp[carclass=="SUV"],pch=c("S"),col=c("purple"))

points(accelrate[carclass=="TS"],msrp[carclass=="TS"],pch=c("T"),col=c("black"))

## Regroup Carclass

cars <- within(cars, {

newclass <- ifelse(carclass == "L", 1, 0)

newclass <- as.factor(newclass)

})

detach(cars)

attach(cars)

## Analysis with Newly Classified Variable (model fit.n)

fit.n <- lm(msrp ~ accelrate + MPG + I(newclass))

fit.nonewclass <- lm(msrp ~ accelrate + MPG)

anova(fit.nonewclass,fit.n)

pairs(msrp ~ accelrate + MPG + I(newclass))

#The predictors seem to be

#related linearly at least approximately.

summary(fit.n)

par(mfrow = c(2,2))

plot(fit.n)

plot(fit.n$fitted.values,StanRes1,ylab="Standardized Residuals")

par(mfrow = c(2,2))

avPlots(fit.n)

#with interaction term

fit.nplus <- lm(msrp ~ accelrate + MPG + I(newclass) + accelrate:MPG)

anova(fit.n, fit.nplus)

summary(fit.nplus) #=> we want to include the interation term

vif(fit.nplus)

par(mfrow = c(2,2))

plot(fit.nplus)

inverseResponsePlot(fit.nplus)

boxcox(msrp~accelrate + MPG + I(newclass) + accelrate:MPG) #log transform

## Model with Log Transform y and Transform x&y

logY <- log(msrp)

logaccelrate <- log(accelrate)

logMPG <- log(MPG)

fit.logY <- lm(logY ~ accelrate + MPG + I(newclass))

fit.log <- lm(logY ~ logaccelrate + logMPG + I(newclass))

#log + interaction term

logY <- log(msrp)

fit.nplus.log <- lm(logY ~ accelrate + MPG + I(newclass) + accelrate:MPG)

par(mfrow = c(2,2))

plot(fit.nplus.log)

AIC(fit.nplus.log)

AIC(fit.logY)

## Compare fit.log and fit.logY and fit.nplus.log

summary(fit.log)

summary(fit.logY)

summary(fit.nplus.log)

fit = fit.nplus.log

#diagnosis plots

par(mfrow = c(2, 2))

plot(fit)

avPlots(fit, layout=c(2,2))

# logY vs. x and Y vs. x

#to see if more linear after the transformation

par(mfrow = c(2, 2))

plot(accelrate,msrp,ylab="msrp")

abline(lsfit(accelrate,msrp))

plot(MPG,msrp,ylab="msrp")

abline(lsfit(MPG,msrp))

plot(newclass,msrp,ylab="msrp")

abline(lsfit(newclass,msrp))

par(mfrow = c(2, 2))

plot(accelrate,logY,ylab="logY")

abline(lsfit(accelrate,logY))

plot(MPG,logY,ylab="logY")

abline(lsfit(MPG,logY))

plot(newclass,logY,ylab="logY")

abline(lsfit(newclass,logY))

#after transformation, obs are more evenly distributed around the line

#values

AIC(fit)

BIC(fit)

# sdres vs. x

par(mfrow = c(2, 2))

StanRes <- rstandard(fit)

plot(accelrate,StanRes,ylab="Standardized Residuals")

abline(h=0)

plot(MPG,StanRes,ylab="Standardized Residuals")

abline(h=0)

plot(newclass,StanRes,ylab="Standardized Residuals")

abline(h=0)

plot(fit.log$fitted.values,StanRes,ylab="Standardized Residuals")

abline(h=0)

#The random nature of these plots is indicative that model fit.logY

#is a valid model for the data

#y vs yhat

par(mfrow = c(1, 1))

plot(fit$fitted.values,logY,xlab="Fitted Values", ylab="log(msrp)")

abline(lsfit(fit$fitted.values,logY))

#shows a plot of Y , price against fitted values, Yhat. We see from

#this figure that Y and Yhat appear to be linearly related. i.e. valid model

#cor

X <- cbind(accelrate,MPG, newclass)

c <- cor(X)

c

#MMP

mmps(fit,layout=c(2,2)) #fit.nplus.log has better mmp

mmps(fit.logY,layout=c(2,2))

par(mfrow = c(1, 1))

mmp(fit,fit$fitted.values,xlab="Fitted Values")

# Assessing Outliers

outlierTest(fit) # Bonferonni p-value for most extreme obs

qqPlot(fit, main="QQ Plot") #qq plot for studentized resid

leveragePlots(fit) # leverage plots

## Ifuential points

cutoff <- 4/((nrow(hybrid)-length(fit.log$coefficients)-2))

plot(fit.logY, which=4, cook.levels=cutoff)

# Influence Plot

#influencePlot(fit.log, id.method="identify", main="Influence Plot",

# sub="Circle size is proportial to Cook's Distance" )

detach(cars)

## - observation 116

no116 <- cars[-c(116), ]

attach(no116)

fit.116.full <- lm(msrp ~., data = no116)

p = 6

step(fit.116.full,direction="both", criterion = "AIC", k = 2\*p)

# BIC

n = length(fit.116.full$residuals)

step(fit.116.full,direction="both", criterion = "BIC", k = log(p))

fit.116.reduced <- lm(formula = msrp ~ accelrate + MPG + carclass, data = no116)

summary(fit.116.reduced)

no116 <- within(no116, {

newclass <- ifelse(carclass == "L", 1, 0)

newclass <- as.factor(newclass)

})

detach(no116)

attach(no116)

fit.116.reduced.new <- lm(formula = log(msrp) ~ accelrate + MPG + newclass, data = no116)

summary(fit.116.reduced.new)

par(mfrow = c(2, 2))

plot(fit.116.reduced.new)

detach(no116)

## Adjusted R Squared

attach(cars)

X <- cbind(accelrate, MPG, newclass, year, MPGMPGe)

b <- regsubsets(as.matrix(X),msrp)

rs <- summary(b)

par(mfrow=c(1,2))

plot(1:length(rs$adjr2),rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")

subsets(b,statistic=c("adjr2"))

detach(cars)

# final plots

par(mfrow=c(2,2))

sj <- bw.SJ(logY,lower = 0.05, upper = 100)

plot(density(logY,bw=sj,kern="gaussian"),type="l",

main="Gaussian kernel density estimate",xlab=expression(logY))

rug(logY)

boxplot(logY,ylab=expression(logY))

qqnorm(logY, ylab = expression(logY))

qqline(logY, lty = 2, col=2)

m2 <- lm(logY~x)

plot(x,ty,ylab=expression(logY))

abline(m2)

summary(m2)

1. A zero-sum game is a mathematical representation of a situation in which each participant's gain or loss of utility is exactly balanced by the losses or gains of the utility of the other participants. [↑](#footnote-ref-0)