SC321 Data Analysis Report 3

Julian Zhu

April 11th 2022

Contents

1 Introduction

Our data comes from information about 110 new car models produced in 2020. We attempt to develop two models to predict the average price of a car. We will also investigate the possible associations between fuel economy and price, and also drive type and price.

2 Variable Information

Our data contains 110 subjects. There are nine variables: 1) Make (the brand of a car), 2) Model (the model of a car), 3) LowPrice (the lowest price of it in 1000's dollars), 4) HighPrice (the highest price of it in 1000's dollars), 5) HwyMPG (highway miles-per-gallon), 6) Seating (seating capacity), 7) Drive (drive type: all-wheel, front-wheel, or rear-wheel), 8) Acc060 (time in seconds it takes to go from 0 to 60 miles-per-hour), 9) Weight (weight in pounds). Variable 1, variable 2, and variable 7 are categorical variables. The remaining ones are quantitative variables. We create a new variable "AvePrice" to document the average of the lowest price and the highest price of each car. This variable is quantitative and measured in 1000's dollars.

3 Methods

We transform AvePrice into log(AvePrice) because the model using AvePrice does not satisfy the conditions for linear regression (see Appendix). The response of our study is log(AvePrice). The predictors that we are interested in are HwyMPG, Seating, Drive, Acc060, and Weight. We use backward elimination of first order linear regression models using these five variables to select the best simple model that can predict log(AvePrice). We then use backward elimination of complete second order linear regression models using the same five variables to select the best complex model that can predict log(AvePrice). We then use the Nested-F test to compare the simple model and the complex model. We use R to perform all analyses.

4 Results

Here is a summary table of descriptive statistics that we are interested in (See Table1).

Variables	Mean	Median	1st, 3rd Quantile	Min, Max	Standard Deviation
HwyMPG	34.00	32.50	28.25, 39.00	21.00, 54.00	7.193269
Seating	5.527	5.000	5.000, 6.500	2.000, 9.000	1.372797
Acc060	7.835	7.700	6.800, 8.675	4.100, 12.100	1.567587
Weight	3875	3845	3292, 4476	2085, 6100	858.8199
AvePrice	43.05	38.29	29.57, 50.23	14.28, 123.35	21.35835
Drive	AWD	FWD	RWD		
Frequencies	80	25	5		

Table 1. Descriptive statistics summary table.

Here is a histograms of the five quantitative variables (See Figure 1).

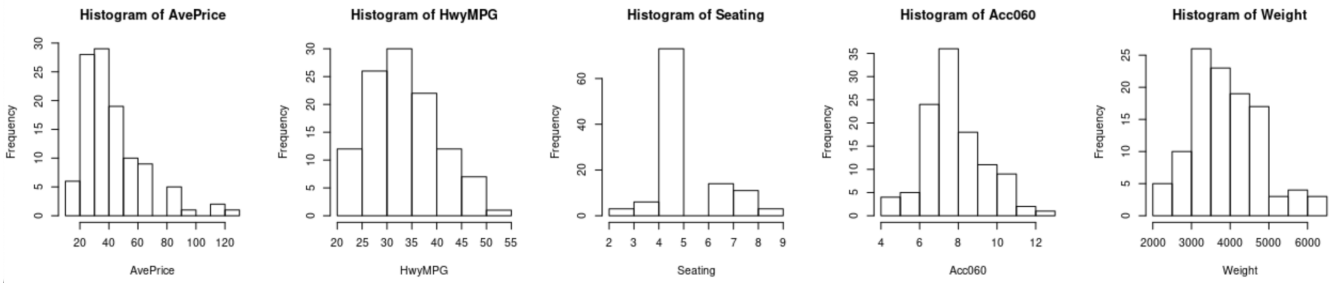

Figure 1. Histograms of five quantitative variables.

AvePrice is heavily skewed right with its median being 38.29 in 1000's dollars. HwyMPG, Seating, and Weight are slightly skewed right with medians being 32.50 miles-per-gallon, 5, and 3845 pounds respectively. There are no 6-seat cars. Acc060 is approximately normal with its median being 7.700. The 95 percent confidence interval for AvePrice is between 0.7126428 and 85.38454.

The simple model is

$$\log(\widehat{AvePrice}) = 3.9374709 - 0.1030556 \cdot \text{Seating} - 0.1259864 \cdot \text{Acc060} + 0.0003407 \cdot \text{Weight} - 0.1529799 \cdot \text{I(DriveFWD)} - 0.1742306 \cdot \text{I(DriveRWD)}. \quad (1)$$

The p-values of all coefficients except the indicator, DriverRWD, are smaller than 0.01 (see Table 2). So there is evidence of associations between the predictors in the model and log(AvePrice). We select this simple model through backward elimination which eliminates the term with HwyMPG (see Appendix).

	Coefficient	Standard Error	t-value	p-value
Intercept	3.937E+00	2.176E-01	18.092	< 2e-16 ***
Seating	-1.031E-01	2.681E-02	-3.843	0.000209 ***
DriveFWD	-1.53E-01	5.629E-02	-2.718	0.007703 **
DriveRWD	-1.742E-01	1.146E-01	-1.520	0.131465
Acc060	-1.26E-01	1.791E-02	-7.035	2.19e-10 ***
Weight	3.407E-04	4.798E-05	7.102	1.58e-10 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.2075 on 104 degrees of freedom				
Multiple R-squared: 0.7977, Adjusted R-squared: 0.7879				
F-statistic: 82 on 5 and 104 DF, p-value: < 2.2e-16				

Table 2. Simple Model Output.

A regression tree (see Figure 2) on all five predictors shows that there might be some interactions between Acc060 and Weight and between HwyMPG and Weight; second order terms with Weight and Acc060 might also be useful.

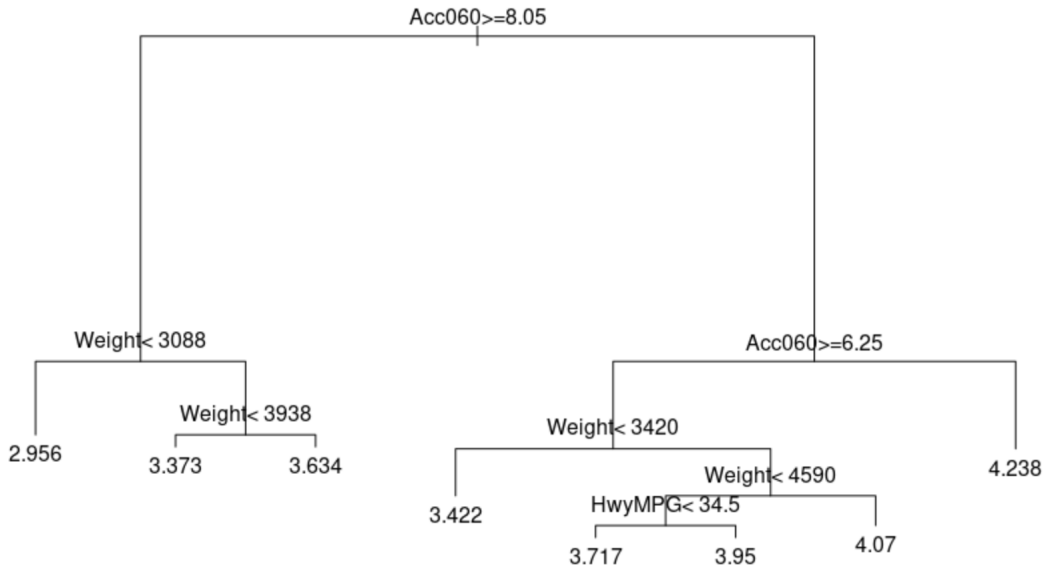

Figure 2. Regression Tree Output.

Therefore, after backward elimination which does not eliminate any terms, our complex model is

$$\begin{aligned}\log(\widehat{AvePrice}) = 9.959 - 0.05124 \cdot \text{HwyMPG} - 0.08167 \cdot \text{Seating} - 0.8331 \cdot \text{Acc060} - 0.001275 \cdot \text{Weight} \\ -0.1555 \cdot \text{I(DriveFWD)} - 0.2977 \cdot \text{I(DriveRWD)} + 0.03475 \cdot \text{Acc060}^2 + 0.0000001035 \cdot \text{Weight}^2 \\ +0.00004214 \cdot \text{Acc060} \cdot \text{Weight} + 0.00001821 \cdot \text{HwyMPG} \cdot \text{Weight}.\end{aligned} \tag{2}$$

The p-values of all coefficients are smaller than 0.1 (see Table 3). So there is evidence of associations between the predictors in the model and log(AvePrice).

	Coefficient	Standard Error	t-value	p-value
Intercept	9.959E+00	1.716E+00	5.805	7.79e-08 ***
HwyMPG	-5.124E-02	2.652E-02	-1.933	0.05616 .
Seating	-8.167E-02	2.484E-02	-3.288	0.00140 **
DriveFWD	-1.555E-01	5.468E-02	-2.844	0.00542 **
DriveRWD	-2.977E-01	1.197E-01	-2.487	0.01454 *
Acc060	-8.331E-01	1.957E-01	-4.258	4.70e-05 ***
Weight	-1.275E-03	5.987E-04	-2.130	0.03567 *
I(Acc060^2)	3.475E-02	7.995E-03	4.347	3.36e-05 ***
I(Weight^2)	1.035E-07	4.457E-08	2.323	0.02222 *
Acc060:Weight	4.214E-05	2.419E-05	1.742	0.08461 .
HwyMPG:Weight	1.821E-05	7.306E-06	2.492	0.01436 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.1842 on 99 degrees of freedom				
Multiple R-squared: 0.8483, Adjusted R-squared: 0.833				
F-statistic: 55.36 on 10 and 99 DF, p-value: < 2.2e-16				

Table 3. Complex Model Output.

The nested F-test comparing the two models give a p-value of 0.00002402 (see Table 4), which is much smaller than 0.001. So we have evidence to support that the complex model has a better fit on our data.

Analysis of Variance Table						
Model 1: simple model						
Model 2: complex model						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	104	4.4796				
2	99	3.3590	5	1.1207	6.606	2.402e-05 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Table 4. Nested F-Test Output.

We care about the possible associations between fuel economy and price and between drive type and price. Given the second model, we provide the 95 percent confidence intervals of coefficients for terms with HwyMPG and Drive in the following table (see Table 5).

Term	2.5%	97.5%
(Intercept)	6.554805e+00	1.336319e+01
HwyMPG	-1.038543e-01	1.371178e-03
HwyMPG:Weight	3.709979e-06	3.270251e-05
DriveFWD	-2.640201E-01	-4.700965E-02
DriveRWD	-5.352324E-01	-6.022163E-02

Table 5. Confidence Intervals for HwyMPG and Drive.

5 Conclusion

We attempt to develop models to predict the average price of a car. We find a complex second-order model with variables HwyMPG, Seating, Acc060, Weight, and Drive the most useful. This model shows that, after controlling for other variables, HwyMPG or fuel economy is associated with price, and the association may vary with different weights. After controlling for other variables, Drive is also associated with price. For future studies, we would like to have a larger dataset to better generalize our results to the entire population. We also recommend studies on possible collinearity between variables not captured by our models.

6 Appendix

1. Condition Check: *Simple Model without Transformation.*

The regression model will be

$$\widehat{AvePrice} = \beta_0 + \beta_1 \cdot \text{Seating} + \beta_2 \cdot \text{Acc060} + \beta_3 \cdot \text{Weight} + \beta_4 \cdot \text{I(DriveFWD)} + \beta_5 \cdot \text{I(DriveRWD)} + \beta_6 \cdot \text{HwyMPG}.$$

Check conditions:

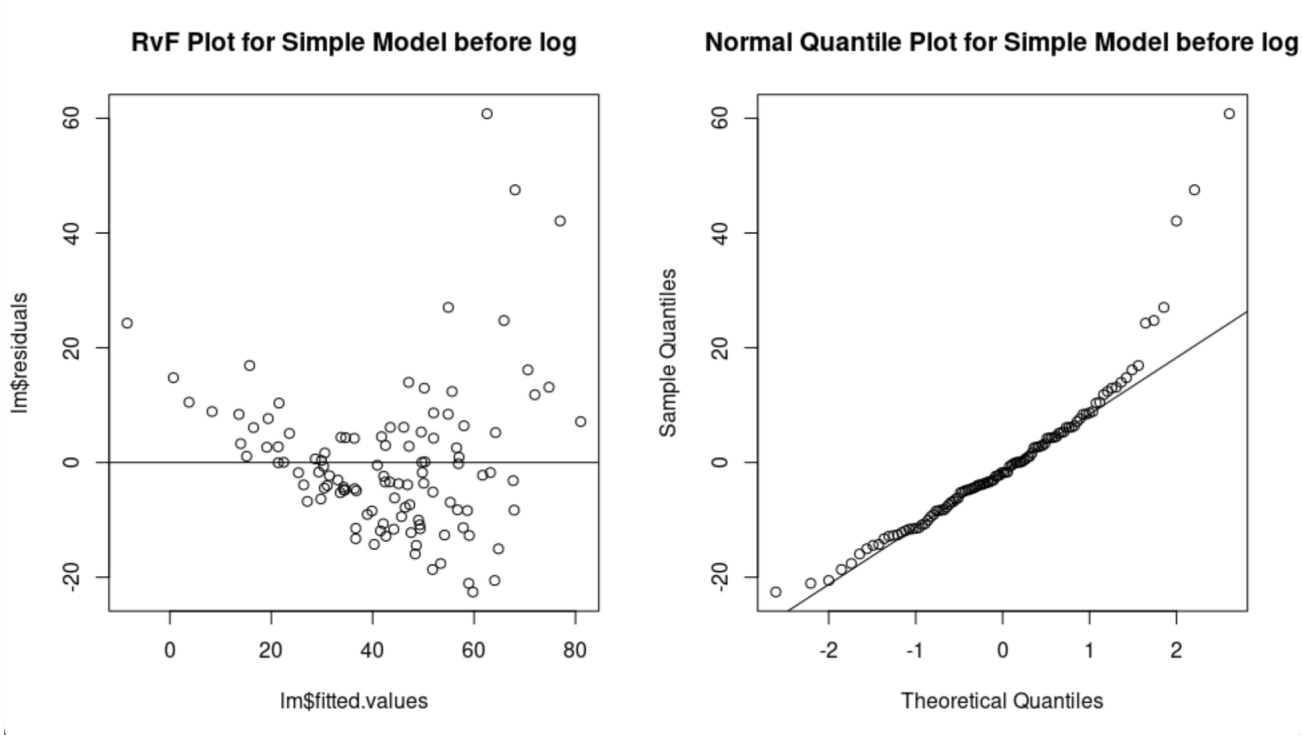

Figure 3. Rvf Plot and NQQ plot.

- For linearity, the RvF plot shows that there are obviously more points clustering under the zero line between 20 and 70. Therefore, it is not appropriate to fit a simple linear regression model.
- For independence, since the observations are collected from distinct individuals, we can reasonably assume that the condition of independence is met.
- For normality, by looking at the normal quantile plot, there is a long tail on the right side of the plot. So we cannot assume that the data is normally distributed.

- For equal variability, since the width of data is much narrower in the middle of the RvF plot, the condition of constant variance is not met.
- For randomness, since we do not know how our data were collected, we cannot assume that our samples are selected randomly. So we cannot generalize the conclusion of our study to the entire population.

2. Condition Check: *Simple Model.*

The regression model will be

$$\log(\widehat{AvePrice}) = \beta_0 + \beta_1 \cdot \text{Seating} + \beta_2 \cdot \text{Acc060} + \beta_3 \cdot \text{Weight} + \beta_4 \cdot \text{I(DriveFWD)} + \beta_5 \cdot \text{I(DriveRWD)}.$$

Check conditions:

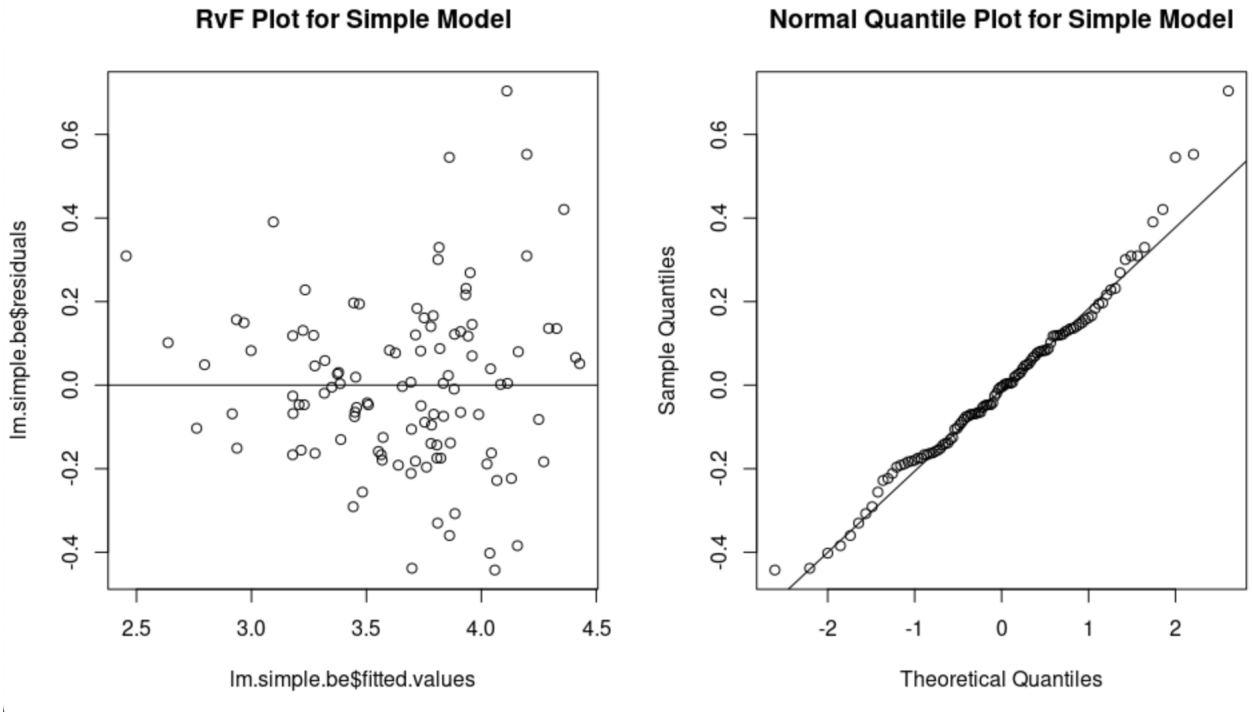

Figure 4. Rvf Plot and NQQ plot.

- For linearity, the RvF plot shows a roughly random distribution of data cases above and below zero in a band of relatively constant width. Therefore, it is appropriate to fit a multiple linear regression model.
- For independence, since the observations are collected from distinct individuals, we can reasonably assume that the condition of independence is met.

- For normality, by looking at the normal quantile plot, we can find that the plot shows a fairly consistent linear trend, and all data cases are very close to fitted line. So the condition of normality is met.
- For equal variability, since the fitted line plot shows that our data spread in roughly equal width (though there might be a slight but acceptable increase) as the fitted value increases, the condition of constant variance is satisfied.
- For randomness, since we do not know how our data were collected, we cannot assume that our samples are selected randomly. So we cannot generalize the conclusion of our study to the entire population.

3. Condition Check: *Complex Model.*

The regression model will be

$$\log(\widehat{AvePrice}) = \beta_0 + \beta_1 \cdot \text{HwyMPG} + \beta_2 \cdot \text{Seating} + \beta_3 \cdot \text{Acc060} + \beta_4 \cdot \text{Weight} + \beta_5 \cdot \text{I(DriveFWD)} + \beta_6 \cdot \text{I(DriveRWD)}$$

$$+ \beta_7 \cdot \text{Acc060}^2 + \beta_8 \cdot \text{Weight}^2 + \beta_9 \cdot \text{Acc060} \cdot \text{Weight} + \beta_10 \cdot \text{HwyMPG} \cdot \text{Weight}.$$

Check conditions:

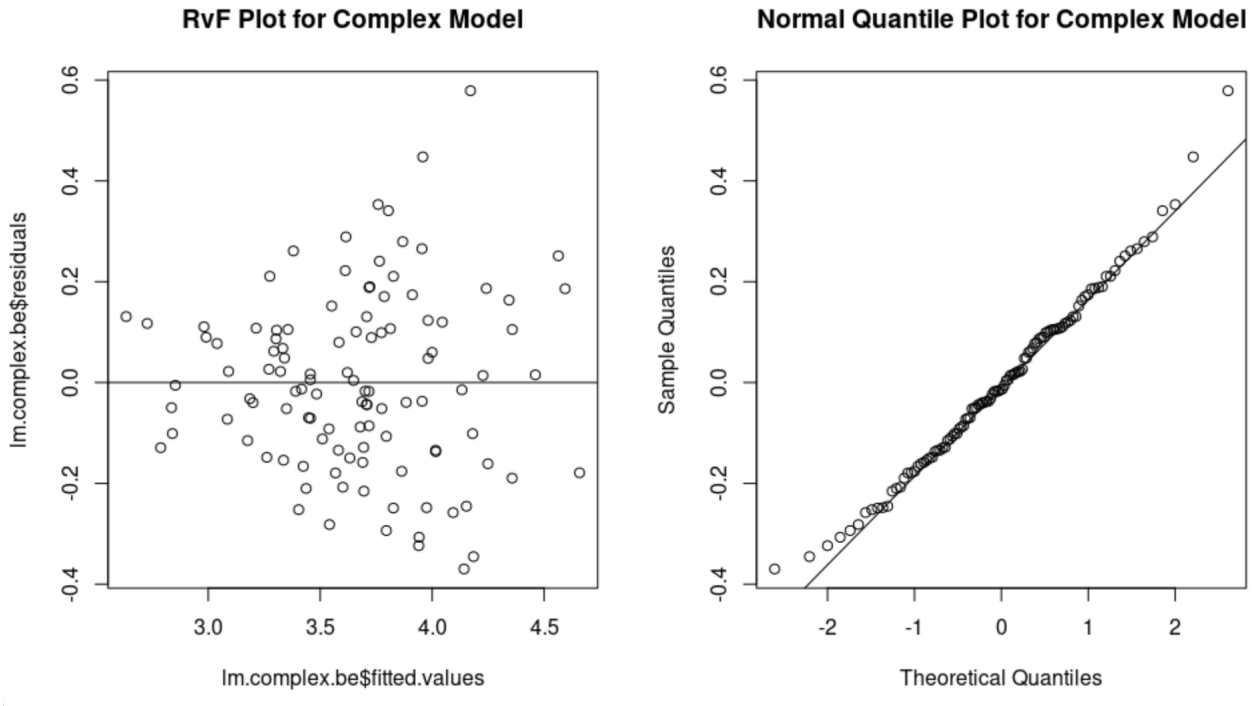

Figure 5. Rvf Plot and NQQ plot.

- For linearity, the RvF plot shows a roughly random distribution of data cases above and below zero in a band of relatively constant width. Therefore, it is appropriate to fit a linear regression model.
- For independence, since the observations are collected from distinct individuals, we can reasonably assume that the condition of independence is met.
- For normality, by looking at the normal quantile plot, we can find that the plot shows a fairly consistent linear trend, and all data cases are very close to fitted line. So the condition of normality is met.
- For equal variability, since the fitted line plot shows that our data spread in roughly equal width (though there might be a slight but acceptable increase) as the fitted value increases, the condition of constant variance is satisfied.
- For randomness, since we do not know how our data were collected, we cannot assume that our samples are selected randomly. So we cannot generalize the conclusion of our study to the entire population.

4. Nested F-Test: *Nested F-test comparing the simple model and the complex model*

The hypothesis are

$$H_0 : \text{the regression coefficients for terms in the simple model is zero;}$$

$$H_a : \text{the regression coefficients for terms in the simple model is not zero.}$$

This is the R output:

Analysis of Variance Table

Model 1: simple model

Model 2: complex model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	104	4.4796				
2	99	3.3590	5	1.1207	6.606	2.402e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The p-value in the R output is smaller than 0.001. So we can reject the null hypothesis and conclude that the complex model fits our data better.

5. R-output Backward Selection: *For Simple Model*

```
Start:  AIC=-339.67
log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 + Weight

           Df Sum of Sq    RSS     AIC
- HwyMPG    1   0.06322 4.4796 -340.10
<none>                  4.4164 -339.67
- Drive     2   0.45889 4.8753 -332.79
- Seating   1   0.62559 5.0420 -327.09
- Weight    1   1.87118 6.2876 -302.81
- Acc060    1   2.12241 6.5388 -298.50

Step:  AIC=-340.1
log(AvePrice) ~ Seating + Drive + Acc060 + Weight

           Df Sum of Sq    RSS     AIC
<none>                  4.4796 -340.10
- Drive     2   0.40806 4.8877 -334.51
- Seating   1   0.63625 5.1159 -327.49
- Acc060    1   2.13167 6.6113 -299.29
- Weight    1   2.17229 6.6519 -298.61

Call:
lm(formula = log(AvePrice) ~ Seating + Drive + Acc060 + Weight)

Coefficients:
(Intercept)      Seating     DriveFWD     DriveRWD       Acc060       Weight
  3.9374709   -0.1030556   -0.1529799   -0.1742306   -0.1259864    0.0003407
```

6. R-output Backward Selection: *For Complex Model*

```
Start:  AIC=-361.77
log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 + Weight +
    I(Acc060^2) + I(Weight^2) + Acc060:Weight + HwyMPG:Weight

                Df Sum of Sq    RSS     AIC
<none>                       3.3590 -361.77
- Acc060:Weight  1   0.10296 3.4619 -360.45
- I(Weight^2)    1   0.18310 3.5421 -357.93
- HwyMPG:Weight  1   0.21071 3.5697 -357.08
- Seating        1   0.36685 3.7258 -352.37
- Drive          2   0.48420 3.8432 -350.96
- I(Acc060^2)    1   0.64099 4.0000 -344.56

Call:
lm(formula = log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 +
    Weight + I(Acc060^2) + I(Weight^2) + Acc060:Weight + HwyMPG:Weight)

Coefficients:
  (Intercept)         HwyMPG        Seating       DriveFWD       DriveRWD
    9.959e+00     -5.124e-02     -8.167e-02     -1.555e-01     -2.977e-01
       Acc060         Weight    I(Acc060^2)    I(Weight^2)  Acc060:Weight
   -8.331e-01     -1.275e-03      3.475e-02      1.035e-07      4.214e-05
HwyMPG:Weight
    1.821e-05
```

R Codes:

```
library(readr)
library(psych)
library(ggplot2)
library(gridExtra)
library(car)
library(rpart)
library(leaps)


# read files
cars <- read_csv("SC321/dar3_cars.csv")
View(cars)
str(cars)
summary(cars)


# re-format variables
cars$Make <- as.factor(cars$Make)
cars$Model <- as.factor(cars$Model)
cars$Drive <- as.factor(cars$Drive)
str(cars)
summary(cars)
summary(cars$Make)

# add AvePrice
r <- nrow(cars)
f <- c()
for (i in 1:r) {
  x <- (cars[i, 3]+cars[i, 4])/2
  f <- rbind(f, x)
}
cars <- cbind(cars, f)
names(cars)[10] <- "AvePrice"
str(cars)
summary(cars)


pairs.panels(cars)
attach(cars)


# Descriptives
mean(AvePrice)
sd(AvePrice)
summary(AvePrice)
UpperQ = mean(AvePrice)+qt(0.975, df = 108)*sd(AvePrice)
LowerQ = mean(AvePrice)-qt(0.975, df = 108)*sd(AvePrice)
UpperQ
LowerQ
```

```
sd(HwyMPG)
sd(Seating)
sd(Acc060)
sd(Weight)

table(Drive)
table(Model)
table(Make)

par(mfrow = c(1,5))
hist(AvePrice)
hist(HwyMPG)
hist(Seating)
hist(Acc060)
hist(Weight)
par(mfrow = c(1,1))

hist(log(AvePrice))



# simple, backward elimination
lm.simple <- lm(log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 + Weight)
summary(lm.simple)
step(lm.simple)
lm.simple.be <- lm(log(AvePrice) ~ Seating + Drive + Acc060 + Weight)
summary(lm.simple.be)
confint(lm.simple.be)
par(mfrow = c(1,2))
plot(lm.simple.be$residuals ~ lm.simple.be$fitted.values, main = "RvF Plot for Simple Model")
abline(0,0)
qqnorm(lm.simple.be$residuals, main= "Normal Quantile Plot for Simple Model")
qqline(lm.simple.be$residuals)
par(mfrow = c(1,1))
# without log, obviously not normal
lm <- lm(AvePrice ~ HwyMPG + Seating + Drive + Acc060 + Weight)
summary(lm)
par(mfrow = c(1,2))
plot(lm$residuals ~ lm$fitted.values, main = "RvF Plot for Simple Model before log")
abline(0,0)
qqnorm(lm$residuals, main= "Normal Quantile Plot for Simple Model before log")
qqline(lm$residuals)
par(mfrow = c(1,1))


# tree
tree <- rpart(log(AvePrice) ~ Seating + Drive + Acc060 + Weight + HwyMPG)
par(xpd = TRUE)
plot(tree)
text(tree, pretty = 0)
```

```
# complex, backward elimination
lm.complex <- lm(log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 + Weight
                 + I(Acc060^2) + I(Weight^2) + Acc060:Weight + HwyMPG:Weight)
summary(lm.complex)
step(lm.complex)
lm.complex.be <- lm(log(AvePrice) ~ HwyMPG + Seating + Drive + Acc060 + Weight + I(Acc060^2)
                    + I(Weight^2) + Acc060:Weight + HwyMPG:Weight)
summary(lm.complex.be)
confint(lm.complex.be)
par(mfrow = c(1,2))
plot(lm.complex.be$residuals ~ lm.complex.be$fitted.values, main = "RvF Plot for Complex Model")
abline(0,0)
qqnorm(lm.complex.be$residuals, main= "Normal Quantile Plot for Complex Model")
qqline(lm.complex.be$residuals)
par(mfrow = c(1,1))


# Nested F
anova(lm.simple.be,lm.complex.be)

detach(cars)
```