**Appendix**

The figures and entries in this appendix are arranged in the order that they appear in the report. R code is displayed for the graphical displays and for any actions that required the use of R. Often R output has been shortened for convenience and ease of interpretation. Some entries provide additional information and evidence for claims that are made in the report.

**Figure 1:** boxplot(Price)

**Figure 2:** boxplot(log(Price))

**Entry 1:**

summary(Price)

Min. 1st Qu. Median Mean 3rd Qu. Max.

8639 14273 18025 21343 26717 70755

> sd(Price)

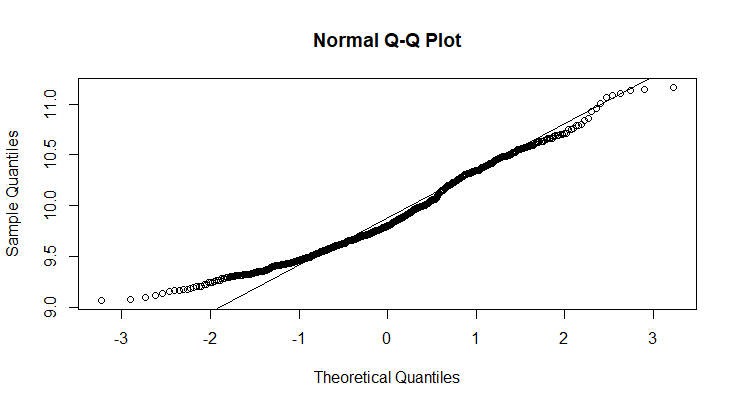
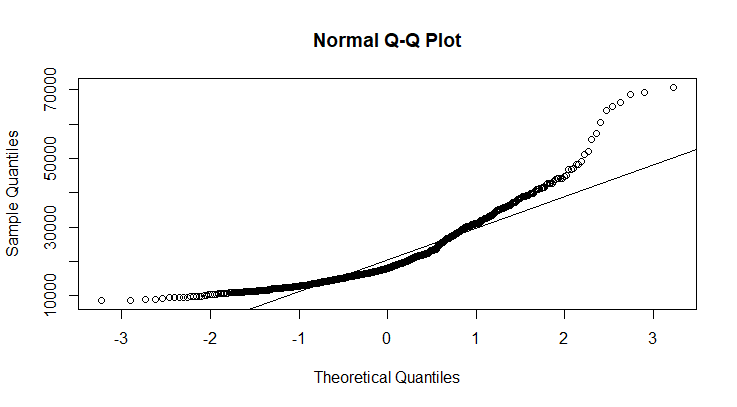
[1] 9884.853

qqnorm(Price)

qqline(Price)

qqnorm(log(Price))

qqline(log(Price))



The summary data for Price shows that the range from the third quartile to the maximum value is $44038, whereas the range from the minimum to the first quartile is $5634, and the IQR is $12444. The Prices are not normally distributed, especially in the upper range, as shown by the QQ plot, but the logarithm transformation does a fairly good job of normalizing the distribution.

**Figure 3:**

plot(Mileage, log(Price), main = "Price vs. Mileage across Make of Car")

points(Mileage[Make=="Buick"], log(Price[Make=="Buick"]),pch=16, col="Red")

points(Mileage[Make=="Cadillac"], log(Price[Make=="Cadillac"]),pch=16, col="Orange")

points(Mileage[Make=="Chevrolet"], log(Price[Make=="Chevrolet"]),pch=16, col="black")

points(Mileage[Make=="Pontiac"], log(Price[Make=="Pontiac"]),pch=16, col="Green")

points(Mileage[Make=="SAAB"], log(Price[Make=="SAAB"]),pch=16, col="Blue")

points(Mileage[Make=="Saturn"], log(Price[Make=="Saturn"]),pch=16, col="Purple")

abline(lm(log(Price[Make=="Buick"]) ~ Mileage[Make=="Buick"]), col="red")

abline(lm(log(Price[Make=="Cadillac"]) ~ Mileage[Make=="Cadillac"]), col="Orange")

abline(lm(log(Price[Make=="Chevrolet"]) ~ Mileage[Make=="Chevrolet"]), col="black")

abline(lm(log(Price[Make=="Pontiac"]) ~ Mileage[Make=="Pontiac"]), col="Green")

abline(lm(log(Price[Make=="SAAB"]) ~ Mileage[Make=="SAAB"]), col="blue")

abline(lm(log(Price[Make=="Saturn"]) ~ Mileage[Make=="Saturn"]), col="Purple")

legend("bottomright", c("Buick", "Cadillac", "Chevrolet", "Pontiac", "SAAB", "Saturn"), pch=c(16, 16),

col=c("red", "orange","black","green","blue","purple"), bty="o",bg="grey",seg.len = 2)

**Figure 4:**

boxplot(log(Price)~Make)

> by(Price, Make, summary)

Make: Buick

Min. 1st Qu. Median Mean 3rd Qu. Max.

14862 18907 21228 20815 22634 26831

----------------------------------------------------------------------

Make: Cadillac

Min. 1st Qu. Median Mean 3rd Qu. Max.

28040 34902 38523 40936 43941 70755

----------------------------------------------------------------------

Make: Chevrolet

Min. 1st Qu. Median Mean 3rd Qu. Max.

8639 12326 14589 16428 17811 47065

----------------------------------------------------------------------

Make: Pontiac

Min. 1st Qu. Median Mean 3rd Qu. Max.

11903 15726 17270 18412 20584 32423

----------------------------------------------------------------------

Make: SAAB

Min. 1st Qu. Median Mean 3rd Qu. Max.

22245 26870 29401 29495 31939 38325

----------------------------------------------------------------------

Make: Saturn

Min. 1st Qu. Median Mean 3rd Qu. Max.

10563 12393 13818 13979 15112 18491

**Figure 5:**

boxplot(log(Price)~Type)

**Entry 2:**

t.test(log(Price)[Sound==0], log(Price)[Sound==1])

Welch Two Sample t-test

data: log(Price)[Sound == 0] and log(Price)[Sound == 1]

t = 3.9128, df = 482.62, p-value = 0.0001043

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.06085982 0.18364171

sample estimates:

mean of x mean of y

9.962071 9.839820

> t.test(log(Price)[Cruise==0], log(Price)[Cruise==1])

Welch Two Sample t-test

data: log(Price)[Cruise == 0] and log(Price)[Cruise == 1]

t = -22.982, df = 743.53, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.5088357 -0.4287475

sample estimates:

mean of x mean of y

9.526290 9.995082

> t.test(log(Price)[Leather==0], log(Price)[Leather==1])

Welch Two Sample t-test

data: log(Price)[Leather == 0] and log(Price)[Leather == 1]

t = -4.2506, df = 540.84, p-value = 2.511e-05

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.17433985 -0.06413294

sample estimates:

mean of x mean of y

9.792737 9.911974

These three pairwise t-tests also produce 95% confidence intervals for the difference in log(price) between having the add-on and not having it. For example, the interpretation of the 95% confidence interval for Leather is this: We are 95% confident that the mean of log(price) is between .064 and .174 lower for a car without leather seats than for a car with leather seats.

**Figure 6:**

boxplot(log(Price)~Sound)

**Entry 3:**

by(log(Price), Sound, summary)

Sound: 0

Min. 1st Qu. Median Mean 3rd Qu. Max.

9.064 9.668 9.935 9.962 10.308 10.927

------------------------------------------------------------------------------------------

Sound: 1

Min. 1st Qu. Median Mean 3rd Qu. Max.

9.091 9.534 9.750 9.840 10.076 11.167

We can calculate the outliers by finding the number of observations that lie above the “upper fence,” which is defined by (3rd quartile – 1st quartile)\*1.5 + 3rd quartile. For cars with an upgraded sound system, the upper fence is at 10.889.

summary(Sound)

0 1

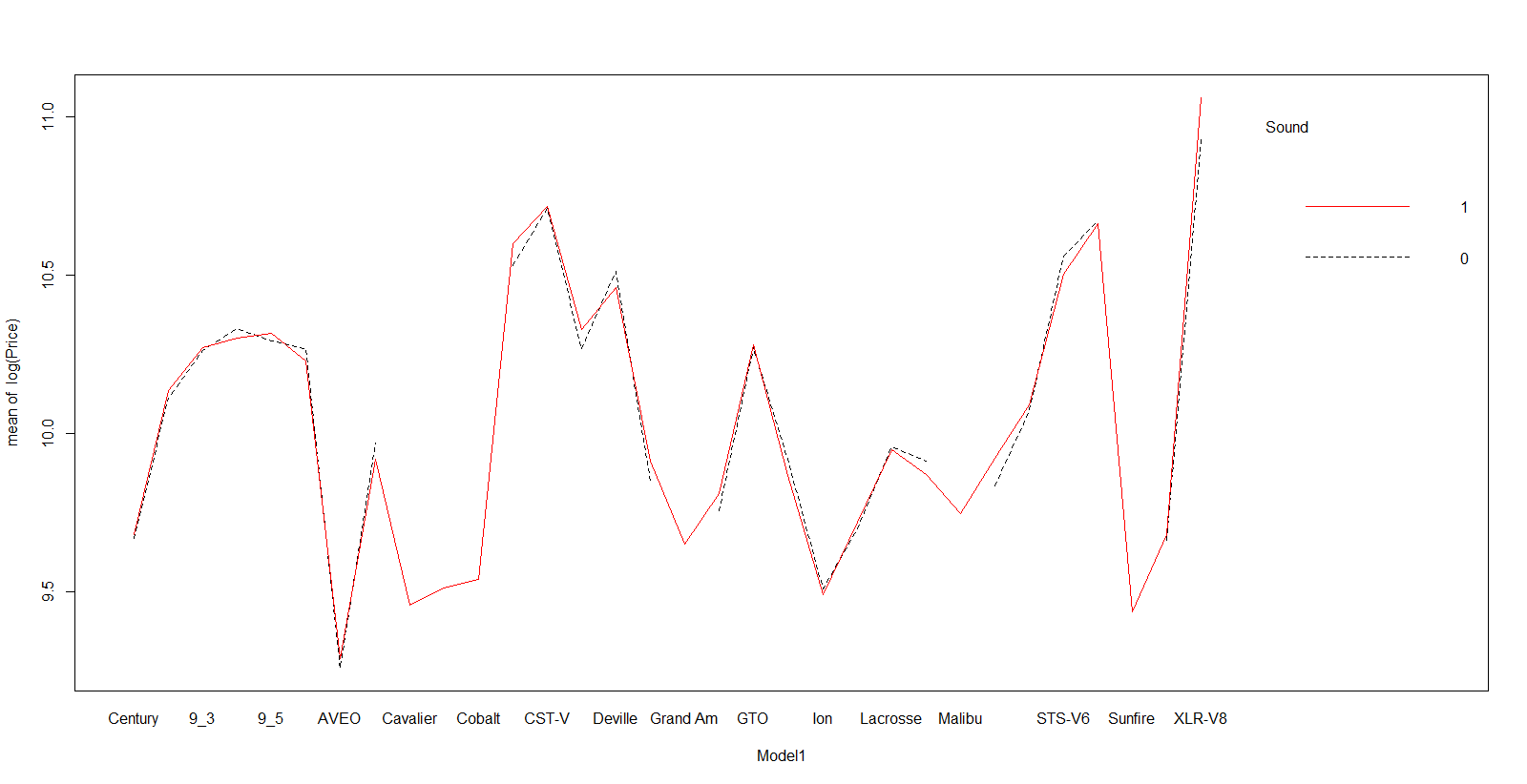
258 546

summary(Sound[log(Price)>10.889])

0 1

1 8

|  |
| --- |
| qqnorm(log(Price)[Sound==1])  > qqline(log(Price)[Sound==1])  > qqnorm(log(Price)[Sound==0])  > qqline(log(Price)[Sound==0])    The distribution for upgraded sound (left) and normal sound (right) are both roughly normal.  Therefore normality is not causing the unusual result.  boxplot(Mileage~Sound)    Upgraded sound systems are distributed evenly along the range of car mileage, so an interaction with  mileage is not causing this trend. |
| **Figure 7:**  interaction.plot(Make1, Sound, log(Price), col=1:6)  **Entry 4:**    interaction.plot(Trim1, Sound, log(Price), col=1:47)  This plot and the following plot show that sound system has basically no effect on price across different  trims and models. |
|  |



interaction.plot(Model1, Sound, log(Price), col=1:32)

**Figure 8:**

interaction.plot(Sound, Leather, log(Price), col=1:2)

interaction.plot(Sound, Cruise, log(Price), col=1:2)

interaction.plot(Cruise, Leather, log(Price), col=1:2)

**Entry 5:** ANOVA model for add-on variables

|  |
| --- |
| aov.fit = aov(log(Price)~(Sound\*Leather)+(Sound\*Cruise)+(Leather\*Cruise))  > summary(aov.fit)  Df Sum Sq Mean Sq F value Pr(>F)  Sound 1 2.62 2.62 22.172 2.94e-06 \*\*\*  Leather 1 3.25 3.25 27.560 1.95e-07 \*\*\*  Cruise 1 32.75 32.75 277.305 < 2e-16 \*\*\*  Sound:Leather 1 1.46 1.46 12.352 0.000465 \*\*\*  Sound:Cruise 1 0.62 0.62 5.259 0.022096 \*  Leather:Cruise 1 0.23 0.23 1.944 0.163601  Residuals 797 94.13 0.12  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | |  | |

Model assumptions:

Independence: Car specifications can reasonably be assumed to be independent of each other.

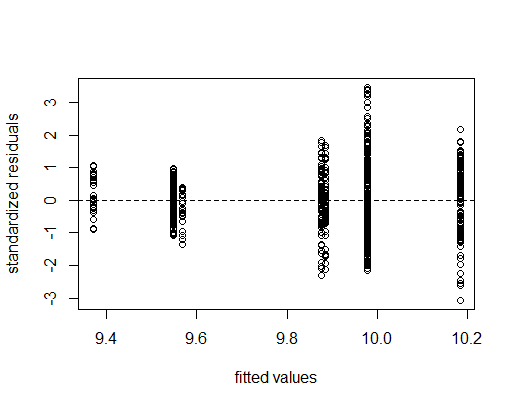
Equal variance:

par(mfrow=c(1,1))

> plot(aov.fit$fitted.values, rstandard(aov.fit),

+ xlab="fitted values", ylab="standardized residuals")

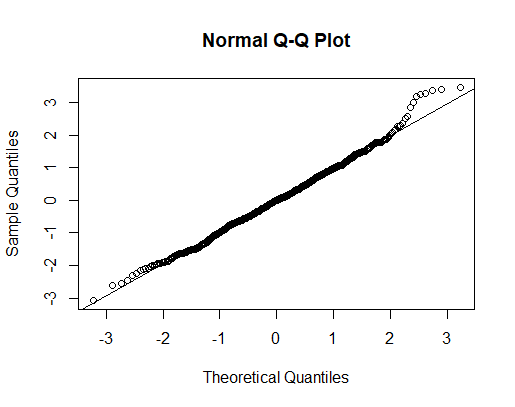
> abline(0,0,lty=2)



Random scatter around 0 shows that the equal variance assumption is satisfied.

Normality:

qqnorm(rstandard(aov.fit)); qqline(rstandard(aov.fit))



The residuals fall along the Q-Q line, so the normality assumption is satisfied.

Additivity Assumption: Based on the interaction plots shown in the main report, this ANOVA model includes interaction terms because the effect of Leather on Sound and Cruise on Sound is not the same at all levels. The p value for the interaction term of Leather\*Cruise is not significant because, as the interaction plot showed, Leather has the same effect on Cruise for both of its levels (leather seats or no leather seats).

**Entry 6:**

fit.add.ons = lm(log(Price)~Sound+Leather+Cruise+(Sound\*Leather)+(Sound\*Cruise))

> summary(fit.add.ons)

Call:

lm(formula = log(Price) ~ Sound + Leather + Cruise + (Sound \*

Leather) + (Sound \* Cruise))

Residuals:

Min 1Q Median 3Q Max

-1.04611 -0.22129 -0.00153 0.23471 1.19418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.33506 0.05306 175.950 < 2e-16 \*\*\*

Sound1 0.13854 0.06787 2.041 0.041540 \*

Leather 0.27694 0.04465 6.203 8.89e-10 \*\*\*

Cruise 0.56333 0.05535 10.177 < 2e-16 \*\*\*

Sound1:Leather -0.19287 0.05717 -3.373 0.000778 \*\*\*

Sound1:Cruise -0.14819 0.06466 -2.292 0.022174 \*

---

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

Residual standard error: 0.3439 on 798 degrees of freedom

Multiple R-squared: 0.3014, Adjusted R-squared: 0.297

F-statistic: 68.85 on 5 and 798 DF, p-value: < 2.2e-16

The model assumptions are satisfied, as shown by the ANOVA model. The linearity assumption is confirmed by the overall p value, which is significant. All covariates are significant at alpha level of 0.05. The interaction term for Cruise\*Leather was found to be insignificant and so was not included in the linear model.

**Entry 7:**

fit.ex = lm(log(Price)~Mileage+Type1+Doors1)

> summary(fit.ex)

Call:

lm(formula = log(Price) ~ Mileage + Type1 + Doors1)

Residuals:

Min 1Q Median 3Q Max

-0.77135 -0.26757 -0.06687 0.24568 0.85111

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.005e+01 3.323e-02 302.363 < 2e-16 \*\*\*

Mileage -8.366e-06 1.487e-06 -5.625 2.57e-08 \*\*\*

Type1Convertible 7.067e-01 5.125e-02 13.788 < 2e-16 \*\*\*

Type1Coupe -1.599e-01 3.307e-02 -4.834 1.61e-06 \*\*\*

Type1Hatchback -3.584e-01 4.721e-02 -7.591 8.84e-14 \*\*\*

Type1Wagon 1.163e-01 4.588e-02 2.534 0.0115 \*

Doors14 NA NA NA NA

---

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

Residual standard error: 0.3451 on 798 degrees of freedom

Multiple R-squared: 0.2964, Adjusted R-squared: 0.292

F-statistic: 67.22 on 5 and 798 DF, p-value: < 2.2e-16

R cannot produce a coefficient for Doors1 because of “singularities,” meaning it is perfectly collinear with another of the variables or combination of the variables in the model. It is simple to see why this is the case: if we know the Type of car, we automatically know the number of doors.

**Entry 8:**

fit3 = lm(log(Price)~Mileage+Model1+Trim1+Cruise+Sound+Leather)

> summary(fit3)

Call:

lm(formula = log(Price) ~ Mileage + Model1 + Trim1 + Cruise +

Sound + Leather)

Residuals:

Min 1Q Median 3Q Max

-0.083117 -0.015251 0.000716 0.015175 0.100862

Coefficients: (8 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.813e+00 8.792e-03 1116.095 < 2e-16 \*\*\*

Mileage -8.149e-06 1.117e-07 -72.960 < 2e-16 \*\*\*

Model19-2X AWD 4.333e-01 1.474e-02 29.391 < 2e-16 \*\*\*

Model19\_3 4.936e-01 1.116e-02 44.231 < 2e-16 \*\*\*

…

Trim1Aero Conv 2D 1.297e-01 1.581e-02 8.205 1.04e-15 \*\*\*

Trim1Aero Sedan 4D -9.893e-02 1.580e-02 -6.260 6.56e-10 \*\*\*

…

Trim1CXS Sedan 4D NA NA NA NA

…

Cruise 5.455e-03 2.799e-03 1.949 0.051697 .

Sound1 6.247e-03 2.199e-03 2.840 0.004630 \*\*

Leather 1.758e-02 2.560e-03 6.868 1.40e-11 \*\*\*

---

Multiple R-squared: 0.9966, Adjusted R-squared: 0.9963

F-statistic: 2973 on 73 and 730 DF, p-value: < 2.2e-16

I have deleted most of the lines of output for the model because there is such a high number of categories for Trim and Model. However, we can still see that there are singularities for certain trims. A brief look at the variance inflation factor for just a few variables reveals another weakness of this model.

vif(fit3)

The R output for this command shows that the VIF for the Chevy Cobalt is 7.2 and the VIF for the Chevy Corvette is 5.7, indicating high multicollinearity.

**Entry 9:**

fit2 = lm(log(Price)~Mileage+Model1+Type1+Cyl1+Cruise+Sound+Leather+(Sound\*Leather)+(Sound\*Cruise))

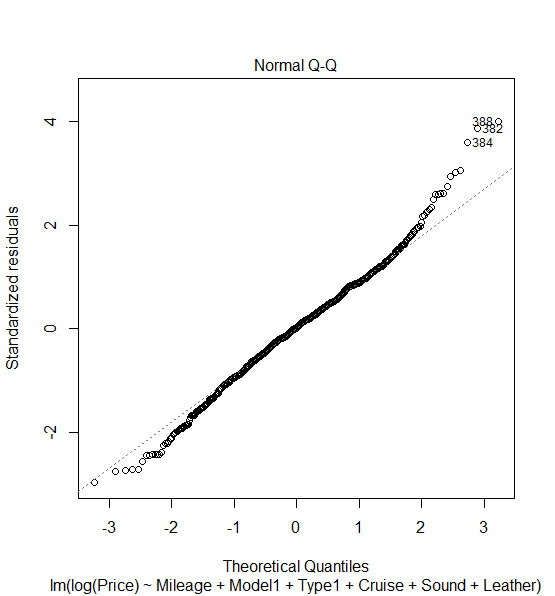
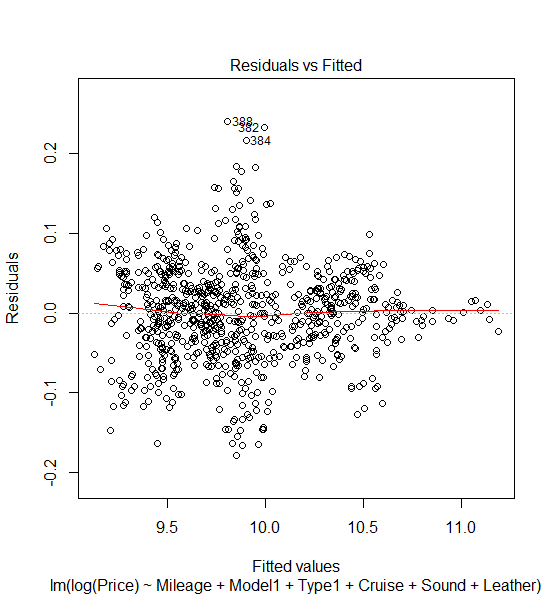
plot(fit2)

Model Assumptions:

Independence: Car specifications can reasonably be assumed to be independent of each other.

Linearity: We saw this was satisfied by the scatterplot of Mileage vs. log(price) at the beginning of the report.

Equal Variance and Normality:



These graphs show that variance is randomly scattered around 0, and residuals are normally distributed, satisfying both assumptions.

vif(fit2)

GVIF Df GVIF^(1/(2\*Df))

Mileage 1.049941 1 1.024666

Model1 66943.703403 31 1.196283

Type1 57.727570 4 1.660248

Cyl1 516.336689 2 4.766869

Sound 8.065218 1 2.839933

Leather 3.201195 1 1.789188

Cruise 5.917363 1 2.432563

Sound:Leather 6.453694 1 2.540412

Sound:Cruise 8.583174 1 2.929705

The VIF for this model shows that multiple variables have a GVIF/(1/2DF)) >2, which indicates multicollinearity.

fit2sel = regsubsets(log(Price)~Mileage+Model1+Type1+Cyl1+Cruise+Sound+Leather+(Sound\*Leather)+(Sound\*Cruise), data = Cars, nbest = 1)

> summaryHH(fit2sel)

model p rsq rss adjr2 cp bic stderr

1 C18 2 0.363 86.0 0.363 23284 -350 0.327

2 C18-Cr 3 0.502 67.3 0.501 18047 -540 0.290

3 Typ1Cn-C18-Cr 4 0.580 56.7 0.579 15080 -671 0.266

4 M1AV-Typ1Cn-C18-Cr 5 0.651 47.2 0.649 12423 -812 0.243

5 M19\_3-M19\_3H-M19\_5-C16-C18 6 0.709 39.3 0.707 10211 -953 0.222

6 M19\_3-M19\_3H-M19\_5-M19\_5H-C16-C18 7 0.790 28.4 0.788 7160 -1207 0.189

7 M19\_3-M19\_3H-M19\_5-M19\_5H-M1STS-V6-C16-C18 8 0.821 24.2 0.820 5978 -1330 0.174

8 Ml-M19\_3-M19\_3H-M19\_5-M19\_5H-M1STS-V6-C16-C18 9 0.848 20.5 0.847 4948 -1457 0.160

Using an all subsets regression, the model with the highest adjusted r-squared, lowest BIC and lowest MSE is a model that contains only some of the different car Models. This result is unrealistic because the regression model needs to be able to predict the price of *every* model of car; this model would be unable to predict the price of, for example, a Chevy AVEO.

backsel = step(fit2, direction="backward")

Start: AIC=-4489.15

log(Price) ~ Mileage + Model1 + Type1 + Cyl1 + Sound + Leather +

(Sound \* Leather) + (Sound \* Cruise)

…

Step: AIC=-4491.7

log(Price) ~ Mileage + Model1 + Type1 + Cyl1 + Sound + Leather

Df Sum of Sq RSS AIC

<none> 2.721 -4491.7

- Sound 1 0.014 2.735 -4489.4

- Leather 1 0.039 2.760 -4482.2

- Cyl1 2 0.158 2.878 -4450.4

- Type1 4 0.915 3.636 -4266.5

- Mileage 1 3.315 6.036 -3853.0

- Model1 31 44.071 46.792 -2266.5

Using backward selection, we remove the variables for Cruise control and the interaction variables for Sound from the model.

**Figure 9:**

hi = hatvalues(fit2)

> plot(hi)

**Entry 10:**

|  |  |
| --- | --- |
| sel = which(hi > 3\*mean(hi))  > Mileage[sel]  [1] 4836 17431 21616 25218 | |
|  | |
|  |  |

For the SAAB 9-2X AWD Linear Wagon 4 Door with 4836 miles on it, the linear model predicts the price as $27,119.27, and the actual price is $27,280. This was calculated using the linear model we identified as the best option.

**Figure 10:**

plot(cooks.distance(fit2), type="h")

> abline(.01,0,lty=2)

**Entry 11:**

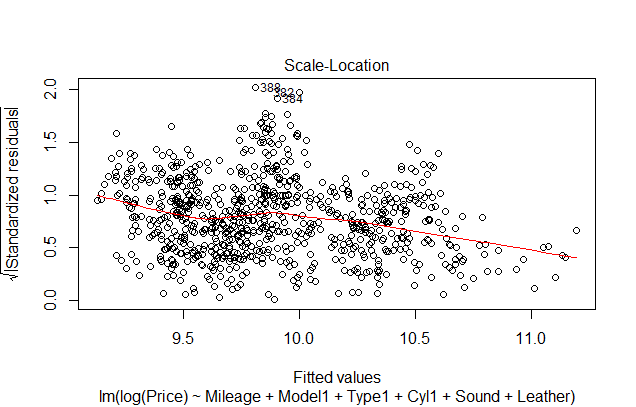
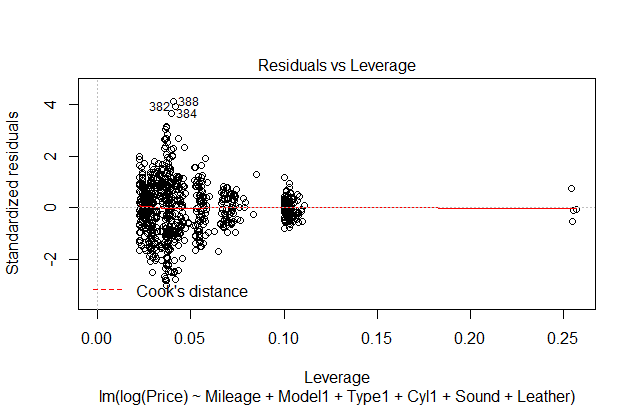
sel2 = which(cooks.distance(fit2)>.01)

Mileage[sel2]

[1] 5379 16111 27940

These are the three most influential points by the criteria of Cook’s distance.

plot(fit2)



This plot of standardized residuals vs fitted values shows that the three observations identified, 382, 384, and 388 have large standardized residuals and are thus influential.

The plot on the right summarizes the findings: there are four high leverage and three high influence points.

**Entry 12:**

Cars.new = Cars[-c(382,384,388,741,741,743,744),]

> Model1new = relevel(as.factor(Cars.new$Model), ref ="Century")

> Type1new = relevel(as.factor(Cars.new$Type), ref ="Sedan")

> Cyl1new = relevel(as.factor(Cars.new$Cyl), ref ="6")

>

> fit2.new = lm(log(Price)~Mileage+Model1new+Type1new+ Cyl1new +Sound+Leather, data = Cars.new)

> summary(fit2.new)

Call:

lm(formula = log(Price) ~ Mileage + Model1new + Type1new + Cyl1new +

Sound + Leather, data = Cars.new)

…

Multiple R-squared: 0.9811, Adjusted R-squared: 0.9801

F-statistic: 983.2 on 40 and 757 DF, p-value: < 2.2e-16

Removing the influential and high leverage observations improves the accuracy of the model.