**Math 240 Final Project:**

Data Analysis of Kelly Blue Book data on 2005 used GM cars

Report Outline:

1. Introduction to dataset
2. Overall Summaries of data set
3. Effect of Add-ons (Sound system, Leather seats, Cruise Control) on price
4. Predictive Models
5. Conclusion

I, **Will Thacher**, pledge that I have not violated the Georgetown University honor code (see http://gervaseprograms.georgetown.edu/honor/). The work I am submitting for this final project is completely my own. I have not communicated with anyone and have not allowed any other student to use or borrow portions of my work. I understand that if I violate this honesty pledge, I will be reported for academic dishonesty to the Honor Council.

Signature : **Will Thacher - 12/18/2017**

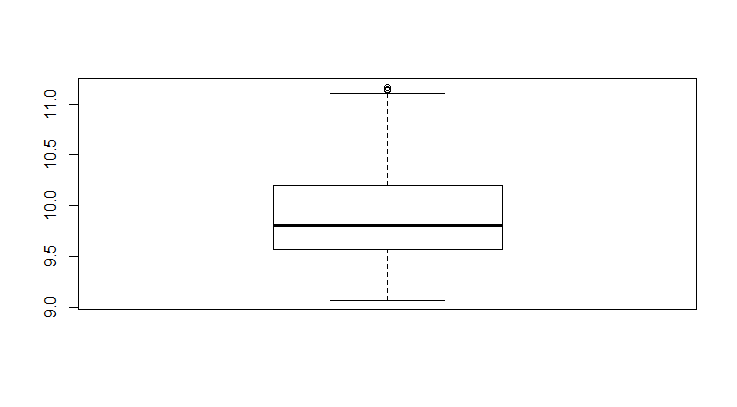
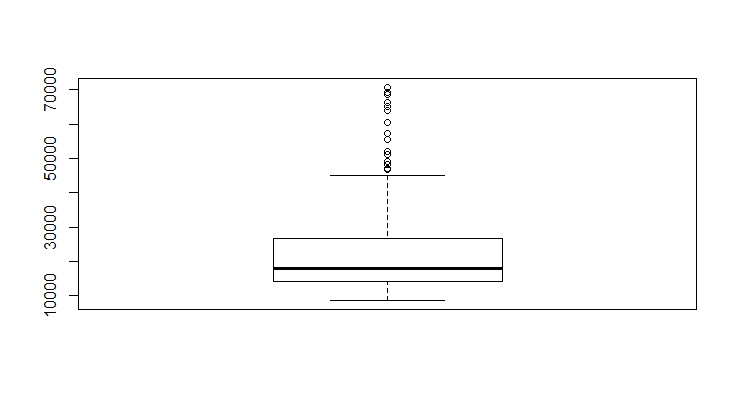
**1. Intro about the model and data**

This dataset from Kelly Blue Book contains prices and vehicle information for six different car brands: Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn. For each car brand, there are different models, and for each model there are different options for trim. For example, the Chevrolet Cobalt comes in four different trims: LS Coupe 2 Door, LS Sedan 4 Door, LT Sedan 4 Door, and Sedan 4 Door. Each combination of make, model and trim corresponds to an engine size (in liters), number of cylinders in the engine, and number of doors. It is important to note that the *model* is unique to the *make*, but the trim can apply across different makes and models. For example, the Buick Century and the Chevrolet Malibu both have the option of a Sedan 4 Door trim.

The vehicle information I have discussed so far (Make, Model, Trim, Cylinders, Liters, Doors) corresponds to *types* of cars, but not to the actual vehicle itself. The dataset also contains information about the Mileage and specific add-ons for each unique vehicle. It notes whether or not the vehicle has an upgraded sound system, leather seats, or cruise control.

Note on transformation of price variable:

I have chosen to use a natural logarithm transformation of the price variable in order to deal with the fact that the range of prices is much larger at higher prices. These graphs illustrate that trend:

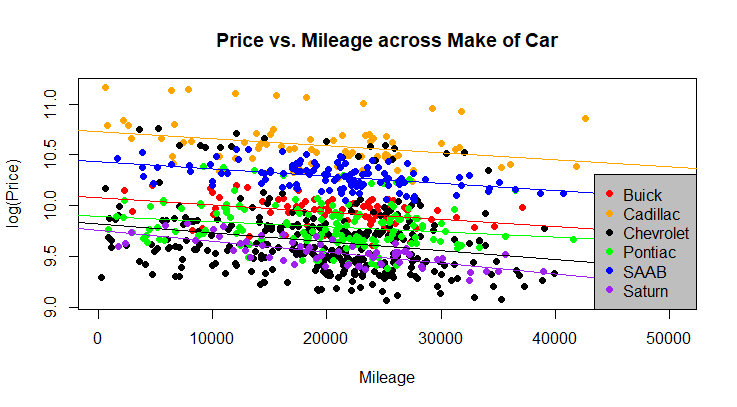


*Figure 1:* Boxplot of Price *Figure 2:* Boxplot of log(Price)

See Appendix entry 1 for further information.

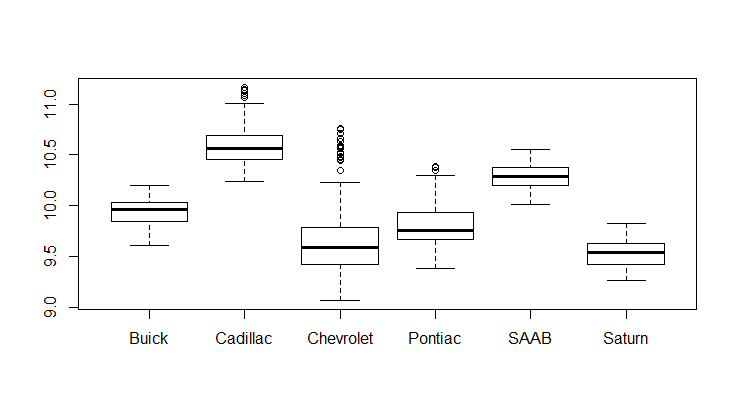
**2. Overall Summaries**

These overall summaries are intended to give an idea of the basic distribution of prices across broad categories.



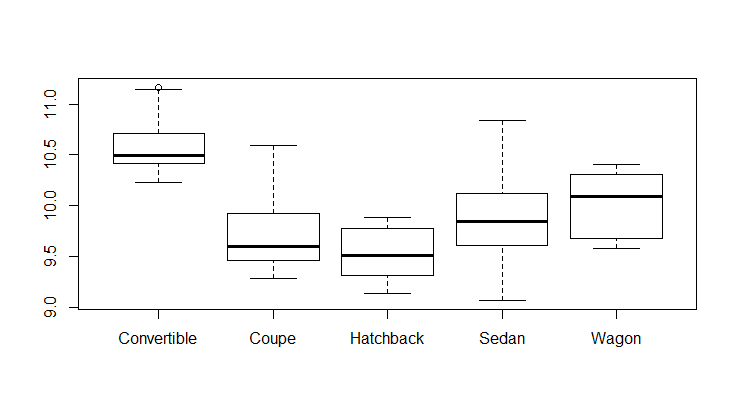
*Figure 3:* Price vs. Mileage across Make of car

This graph shows several important generalizations for this dataset. As would be expected, Mileage is negatively correlated with Price across all Makes. This graph also gives an idea of the average prices for cars of various Makes, with Cadillac being the highest and Saturn the lowest. The slopes of the lines are close to parallel, meaning that Mileage affects price in a similar way across Make of car.



*Figure 4:* Boxplot of log(Price) across Make of car

This boxplot, as well as the summary statistics in the Appendix under Figure 4, show that Chevrolet, Pontiac, and Cadillac have the widest range of prices. The range from the third to the maximum price for Cadillac and Chevrolet are quite large; they are $26,814 and $29,254, respectively. Those ranges are at least four times larger than those for Buick, SAAB, and Saturn.



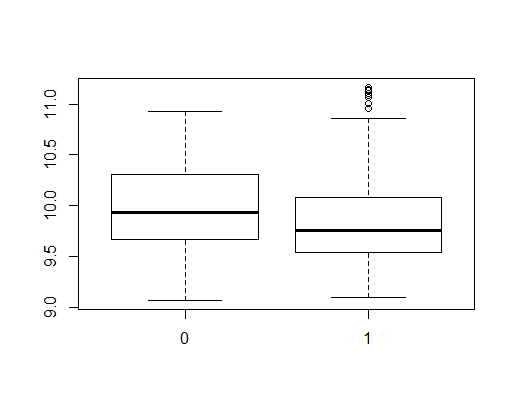
*Figure 5:* Boxplot of log(Price) across Type of car

This boxplot shows that Convertibles are the most expensive type of car on average and that Coupes and Sedans have the largest range of prices. The IQRs of Coupes, Hatchbacks, and Sedans have quite a bit of overlap, and the IQR of Wagons has a little bit of overlap as well.

**3. Effect of Add-ons on Price**

This section will examine the effect of “add-ons” (upgraded sound system, cruise control, and leather seats) on car price. Pairwise t-tests confirm that the mean of log(price) is different for upgraded sound system or not, cruise control or not, and for leather seats or not. See Appendix Entry 2.

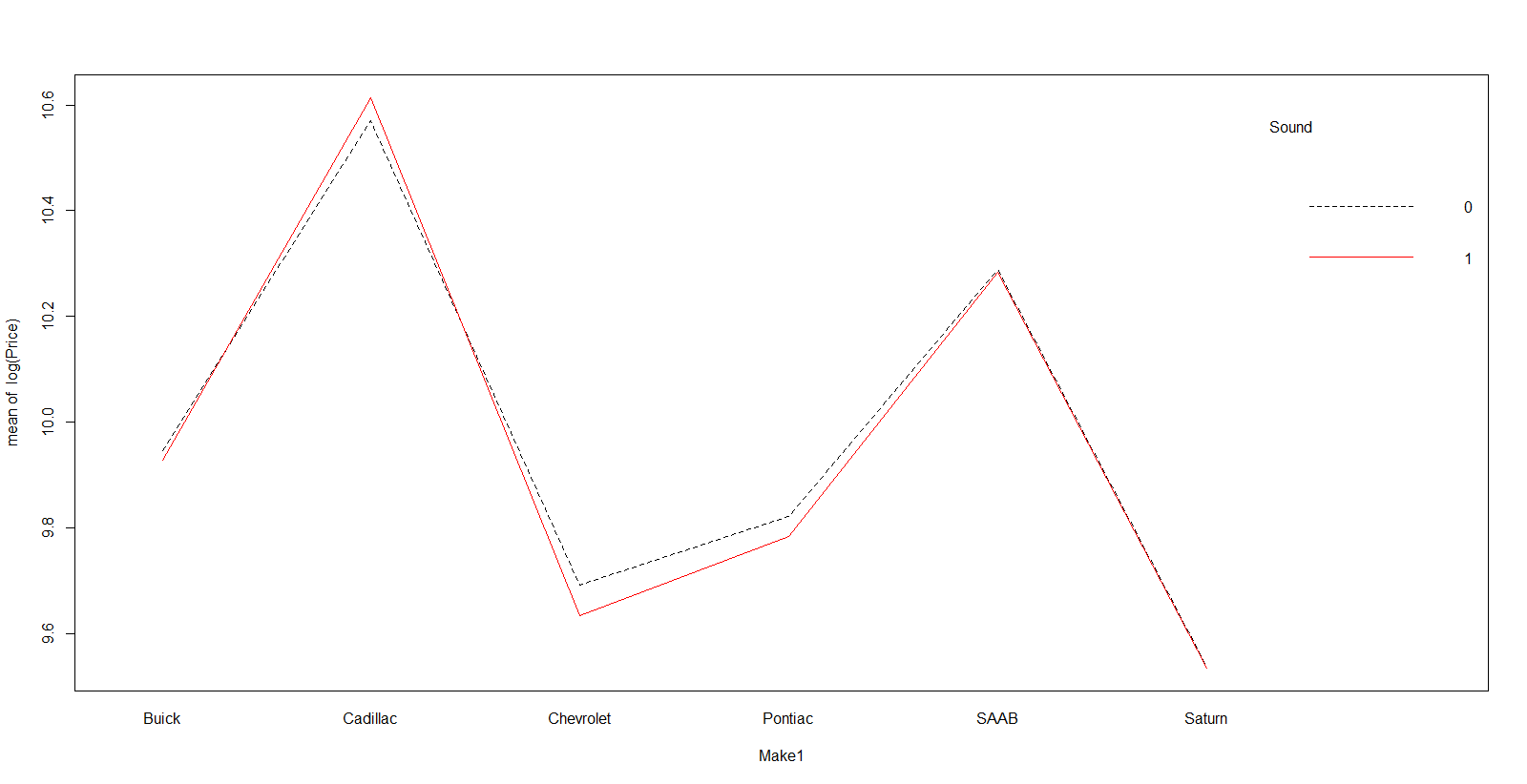
As would be expected, the price for a car with cruise control is higher than for a car without cruise control, and the price of a car with leather seats is higher than for a car without leather seats. However, the pairwise t-test for the sound system tells us that we are 95% confident that the mean of log(price) is higher for a car *without* an upgraded sound system. This surprising result requires closer examination:



*Figure 6:* Boxplot of log(Price) for normal (0) and upgraded (1) sound systems. See Appendix Entry 3 for further information.

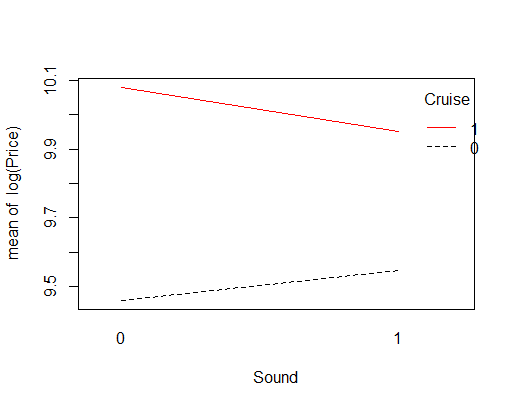
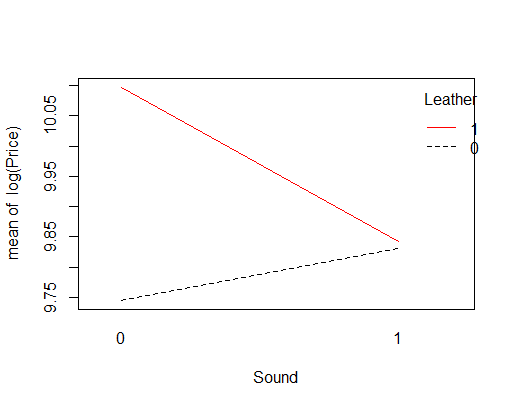
There are only 8 outliers for upgraded sound systems, less than 2% of the total for cars with upgraded sound systems, so outliers are not causing this result. The distribution for log(price) for upgraded sound systems is skewed slightly right, but not severely so. The range of prices for either sound system option is roughly equal. Perhaps interaction effects are causing this unexpected phenomenon; it is possible that another variable that is closely associated with upgraded sound systems causes is itself associated with lower car prices.

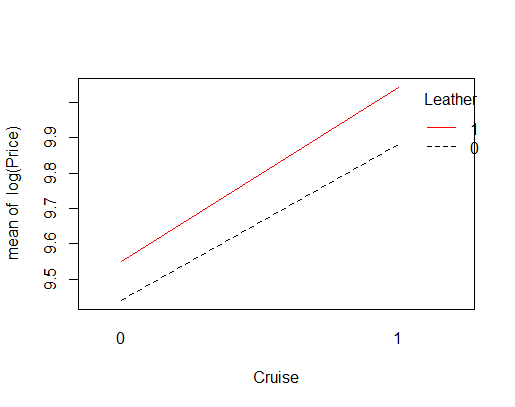
After examining interaction plots for log(price) between Sound and other variables in the model, the only variables interacting unusually with Sound were the variables for cruise control and leather seats. Here is an example of an interaction plot that shows no interaction:



*Figure 7:* Interaction plot of mean of log(price) for Make of car and Sound system. See Appendix Entry 4 for further information.

This plot, as well as those in the Appendix, show that sound system has basically no effect on the price of the car across categories of Model and Trim. The question remains: why do cars *without* upgraded sound systems generally have higher prices? The interaction plots between the add-on variables are informative:





*Figure 8:* Interaction plots for all combinations of add-on variables for mean of log(price)

The interaction plots between Sound and Leather, and Sound and Cruise show that car price is indeed higher for cars with upgraded sound systems if the car does not have another add-on. However, the car price is *lower* if the car has an upgraded sound system and another add-on. We would expect that all three of these interaction plots would look like the one between Leather and Cruise, which shows that add-ons increase price uniformly.

Using multiway ANOVA, we can determine that the mean of log(price) is different for different combinations of add-ons. See Appendix Entry 5 for ANOVA model. Fitting a linear model for the add-ons with the addition of interaction terms gives the following result:

**log(Price) = 9.34 + .14Sound + .28Leather + .56Cruise - .19(Sound\*Leather) - .15(Sound\*Cruise))**

See Appendix Entry 6 for further information on this model.

If we consider a car that only has one of these add-ons, all three increase the price of the car, on average, with cruise control having the greatest effect and sound system having the least effect. Despite the fact that the coefficients for the two interaction terms are negative, the average price for a car with any combination of add-ons is higher than the price for a car without add-ons or with any single add-on. The average of log(price) for different combinations of add-ons is summarized in this table:

Average log(Price) Add-ons

|  |  |
| --- | --- |
| 9.34 | None |
| 9.48 | Sound |
| 9.62 | Leather |
| 9.90 | Cruise |
| 9.57 | Sound +Leather |
| 9.89 | Sound + Cruise |
| 10.18 | Leather + Cruise |
| 9.98 | Sound+Leather+Cruise |

In concluding this section, we find that the unusual trend in price related to sound system is explained somewhat by interaction terms with the other add-on variables. We can tentatively conclude that car buyers do not value upgraded sound systems as much as they value leather seats and/or the cruise control function.

**4. Predictive Models**

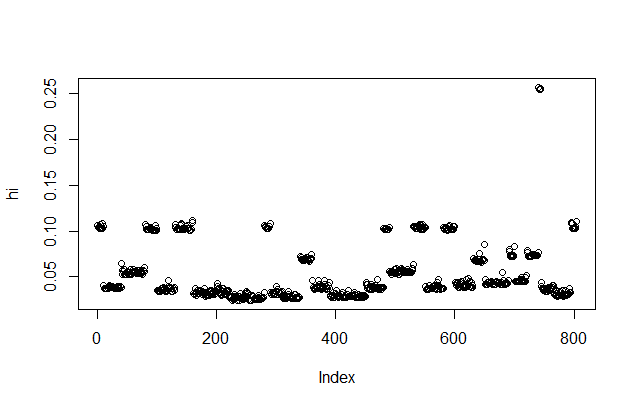
An important consideration when creating this model is what it will be used for. Kelly Blue Book provides two basic services: information to buyers about how much others payed in the same area for the same car and information to sellers about how much their vehicle is worth. While the first service relies on historical information, the second one utilizes predictive modeling. We must identify the model that gives us the best price prediction.

Some of the pairs of covariates cannot be included in the same model because they are perfectly collinear. For example, if we include type of car in the model, including number of doors adds nothing to the model because a Sedan, a Wagon, or a Hatchback always has 4 doors and a Convertible or a Coupe always has 2 doors. Another example is that if we have a regression model with Model and Trim as covariates, adding the number of cylinders and the engine size in liters to the model adds nothing; every combination of model and trim stipulates a certain engine size and number of cylinders. See Appendix Entry 7.

The first model we will examine has the covariates Mileage, Model, Trim, Sound, Cruise, and Leather. It does not include Make, Liters, or Cylinders, because that information is already provided by the combination of Model and Trim. This model has an adjusted R-squared of .996, meaning it is a very good fit. However, the model contains many instances of multicollinearity. This is expected, given that some models and trims always come with certain add-ons; for example, the SAAB 9\_5 Linear Wagon 4D always comes with cruise control, so this specific trim can be written as a linear combination of Cruise and Model. Therefore, we must use a model with either only Trim or only Model. See Appendix Entry 8 for further information.

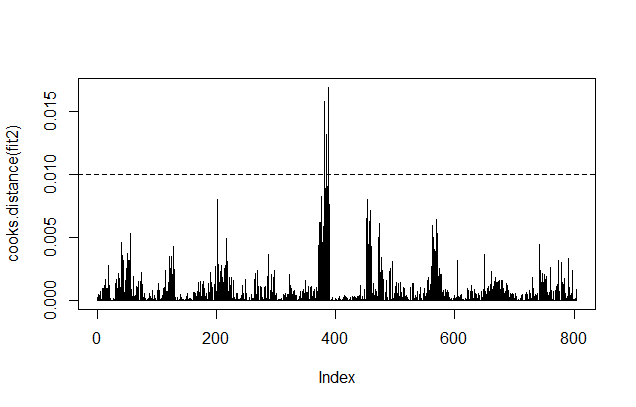
The next model we will look at uses the covariates Mileage, Model, Type of car, number of Cylinders, Cruise, Sound, Leather, and interaction variables for Sound. We have included these interaction variables because we saw in the previous section that Sound affects price differently for different combinations of add-ons. This model also displays high multicollinearity, as shown in Appendix Entry 9. The addition of Liters to the model added nothing because of its prefect multicollinearity with other variables. Using backwards selection, we determined that the variable Cruise, as well as the interaction variables for Sound, should be removed to obtain a more effective model.

We can examine this model to see if there are any data points with high leverage or high influence:



*Figure 9:* Plot of leverage for model with covariates Mileage, Model, Type of car, number of Cylinders, Sound, and Leather. See Appendix Entry 10.

There are four high leverage observations, and they all correspond to the SAAB 9-2X AWD Linear Wagon 4 Door. There are only four observations of this specific model and trim combination, so that is most likely why they have high leverage. The model is actually quite accurate at predicting the price of these vehicles that correspond to high leverage observations.



*Figure 10:* Cook’s distance for all observations to indicate influence of points. See Appendix Entry 11 for further information.

The three most influential points correspond to three Chevy Impala SS Sedan 4D models. These points correspond to vehicles that have only one or two add-ons, whereas most of the other Impalas SS Sedan 4Ds have all three add-ons. If we remove these influential and high leverage points from our model, we raise the adjusted R-squared by .001 to .98. See Entry 12.

**6. Conclusion**

In this report we have looked at an overview of the data and have seen how car price differs across different categories, and we have analyzed how add-ons affect car price. We found that a car’s sound system is not as important as whether it has leather seats or cruise control in determining price. We have also created a linear model using the covariates Mileage, Model, Type of car, number of Cylinders, Sound, and Leather. This model has strengths and weaknesses: it has a very high adjusted R-squared and a low AIC, but it also displays multicollinearity. This multicollinearity is virtually unavoidable for this dataset given that certain car specifications, which in this model correspond to categorical variables, often are found in certain combinations. Despite this obvious downside, the model still has the ability to perform one of Kelly Blue Book’s main services: outputting an accurate car price based on specifications for one specific car.