Case Study: Cars

Stephen Merrill
Kristina Murri

February 11, 2017

Abstract

The price at which a used car will sell is a matter of debate between the salesman and the customer. Often, used car dealers receive a variety of cars into their lots, and then they must determine the price at which they will offer to sell the vehicle. Since this may be a complicated process, we use regression from data on used car sells in order to find out which factors best aid in predicting the price for which a used car will sell. As there are non-linearities in the data and with some of the explanatory variables, a generalized additive is used with a natural spline on one of the variables. Best subset selection with cross-validation is used in order to determine which covariates should be used. After cross-validating the model, we find that the model fits the data fairly well and has good predictive accuracy. The selling price is determined mostly by the manufacturing year, weight, number of miles, horsepower, manufacture, automatic air conditioning, and powered windows. Therefore, both internal and external features are important in the future selling price of any given used car.

1 Purpose

Used car dealerships seek to swindle their customers by purchasing a used car at a low price and then reselling the car at a much higher price. If the dealerships know the value a car will sell for, they will hold an advantage over their customer.

Therefore, in order to aid these in-famous car salesmen, our purpose is to create a model that will predict the future selling price of a used car given characteristics of the car.

Although we have a data set of 1436 observations and 21 descriptive explanatory variables about the characteristics of the cars, this data set contains several problems. Most noteably is the clear non-linear relationship between Miles and Price. In order to build a model, this non-linearity needs to be addressed in some manner. There are also apparent outliers in the data. One, in the cc data, was a misprint and was easily corrected. However, others in the Horsepower data were not as easily rectified and will be considered in the results. Finally, some data cleanup was necessary. The cylinder data was found to be uninformative and thus removed. Additionally, we were considered about sparse observations in categories in the Color, Doors, and Air Conditioning data. However, none of the data combination techniques we explored seemed to provide extra accurateness or effectiveness, so we left this data as is. A brief glance at the data can be seen in Figure 1.

The goal of this analysis is to predict the future used car selling price, given characteristics about the used car.

2 Generalized Additive Model (GAM)

Since there are variables with a nonlinear relationship with respect to the selling price, we need to account for them using some type of linear regresssion. We choose to use natural splines to account for these variables, because we can still use parametric functions to interpret them as well as they predict well on the tails of the distributions.

Then, we use a Generalized Additive Model (GAM), because it is an extension of the standard linear model but additionally allows for the use of non-linear functions, such as we have with this data set.

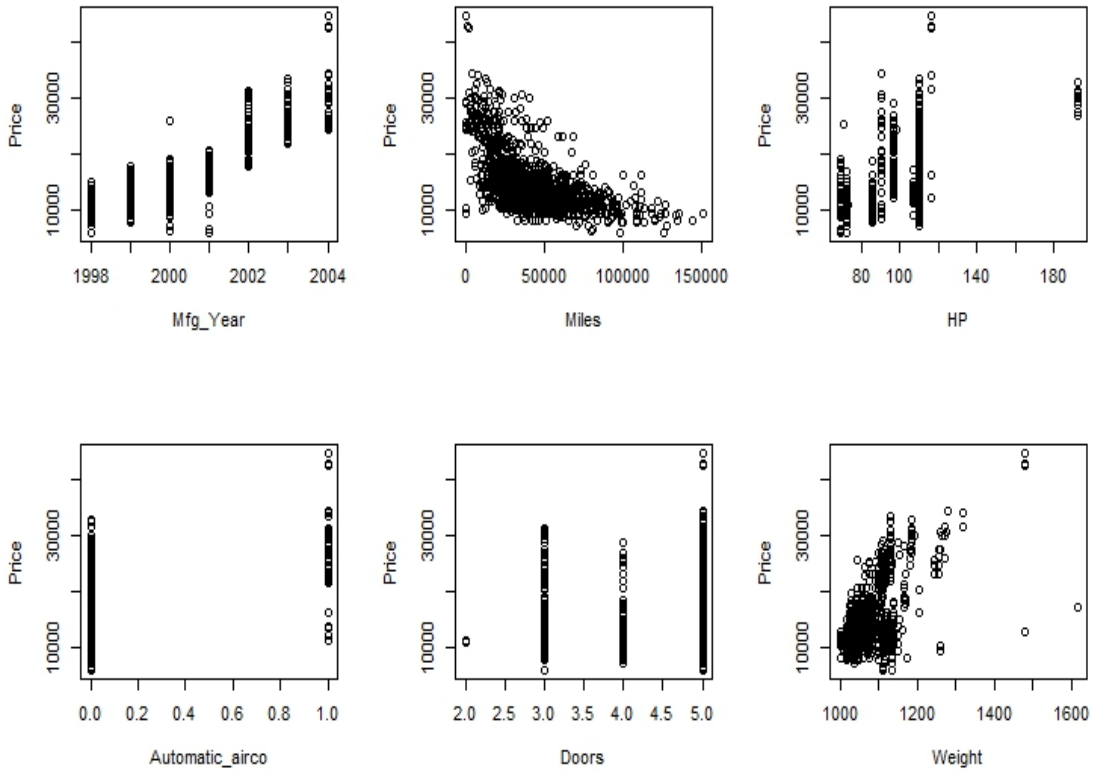


Figure 1: Scatterplot matrix of selected covariates and price

Our GAM model is defined by functions for each variable:

$$y_i = \_0 + \sum_{p=1}^{P} f_p(x_p) + \epsilon_i \tag{1}$$

$\epsilon_i \sim N(0, \sigma^2 I)$

Where:

f_1 is a Natural spline for the Miles variable with three degrees of freedom and two knots. $f_2 \dots f_{P=6}$ are each linear functions with respect to the explanatory variables of Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows. The generalized additive model includes an overall intercept and an overall error term as well as a coefficient measuring the slope for each of the explanatory variables.

The B_0 represents the *Price* when all of the predictor variables are equal to zero. Each of the B_i in the linear functions of the Generalized Additive Model are representations of slope, meaning they are the amount that the Price will increase by for each one unit increase in each explanatory variable, respectively.

Our model allows for both inference and prediction to be made. For this problem, interest lies in making predictions of used car sales prices. The model accomplishes this goal by determining the effect sizes (the β values) of the significant variables outlined above, as well as the relationship between Price and Miles. Once determined, prediction for balance can be made by gathering data on the cars and making calculations according to the model.

2.1 Model Assumptions

Since we are using a Generalized Additive Model, the model assumptions are the same as those for a linear model.

Linearity Each variable must have a linear relationship with the response. If this is not the case, the entire model is invalid since it would be fitting a line to non-linear data. However, we know that Miles is non-linear with Price, and we've fit a natural spline accordingly. Therefore, this assumption only needs to hold for the linear functions in the GAM.

Independence The data must be independent. If this assumption is violated, measures of variability will typically be too small. This is a difficult assumption to verify. Usually prior knowledge of the data is required.

Normality The errors must be normally distributed. Otherwise, confidence and prediction intervals that depend on t distributions are incorrect.

Equal Variance The errors also must have equal variances. Without this, measures of variability will once again be invalid.

3 Model Justification and Performance

We used two Cross-Validation techniques in order to build the model. For each technique, we randomly generated 100 different test and training sets from the original data set in order to have consistent results.

We first used Cross-Validation to determine the optimal number of degrees of freedom to use in the natural spline for Miles. This was determined to be three. Results can be seen in Figure 2.

We then used best subset selection in order to select the functions for our GAM model. This process selected nine variables - the natural spline, which counted as three due to its degrees of freedom, and the other six variables, which we considered linear functions in order to add them to the GAM. We therefore fit the GAM with seven functions. Results can be seen in Figure 3.

3.1 Model Assumptions Verification

Linearity The scatterplot matrix (Figure 4) shows a vague linear relationship between each of the nine explanatory variables and the response. No other kind of relationship is prominent, so we considered this assumption to be met. As previously discussed, Miles is not considered.

Independence Independence is difficult to determine. Because there is no prior knowledge that would suggest a violation of this assumption, it is assumed to be met.

Normality This histogram of the residuals (Figure 5) show that the errors are skewed and not precisely normally distributed. This means that our interval estimates that depend on t distributions will be inaccurate.This also implies that the standard errors will be inflated.

Equal Variance This plot of the residuals (Figure 6) offers no evidence of an unequal variance.

3.2 Model Fit and Prediction

The model fits the data as determined by the minimization of RMSE during Cross-Validation techniques, and R^2, which had a value of .8959. This means that 89% of the variability in Price is explained by the variability in all of the covariates.

The goal of this study was to predict the price of used cars. Using Cross-Validated testing and training sets, we simulated prediction for many different test sets and recorded the results in Table 1 and Figure 7.

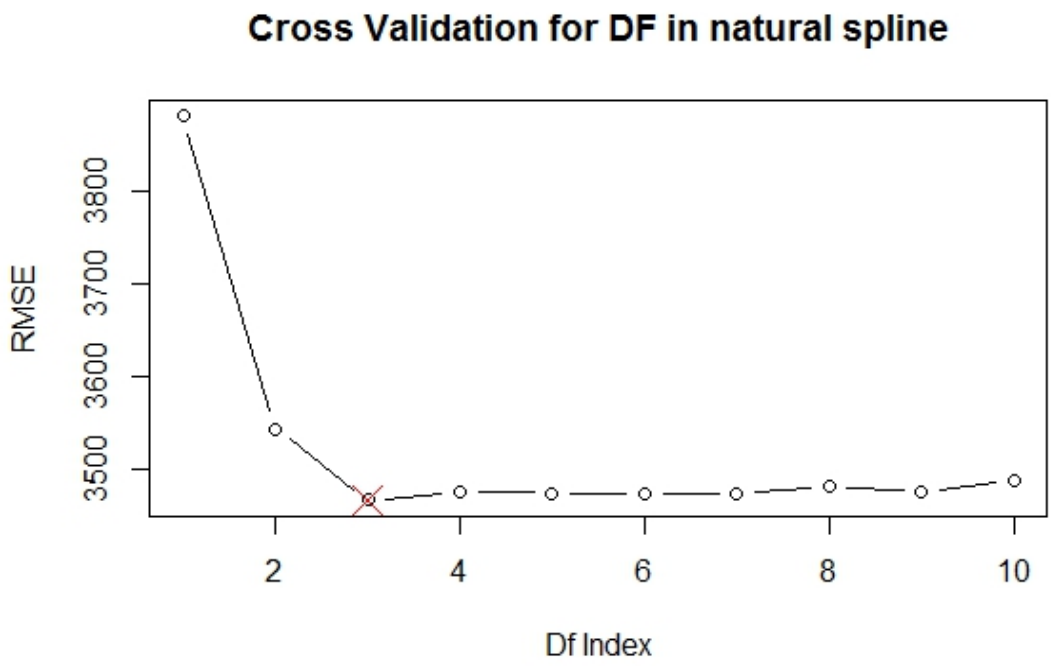


Figure 2: Three degrees of freedom in the natural spline minimized RMSE

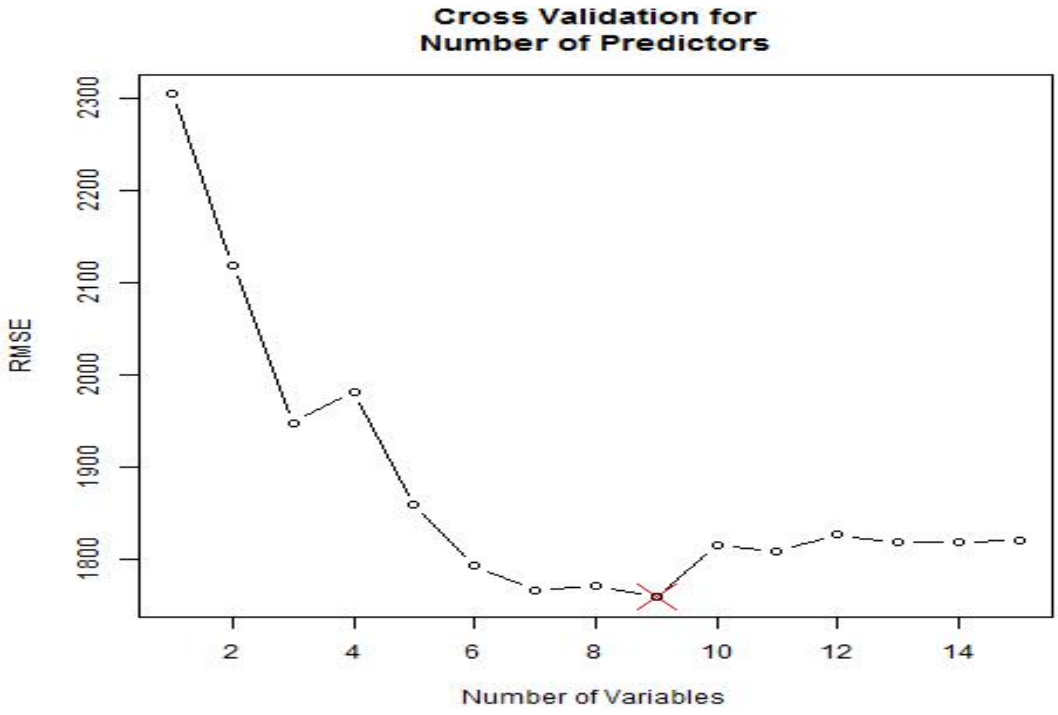


Figure 3: Best subset indicated nine variables minimized RMSE

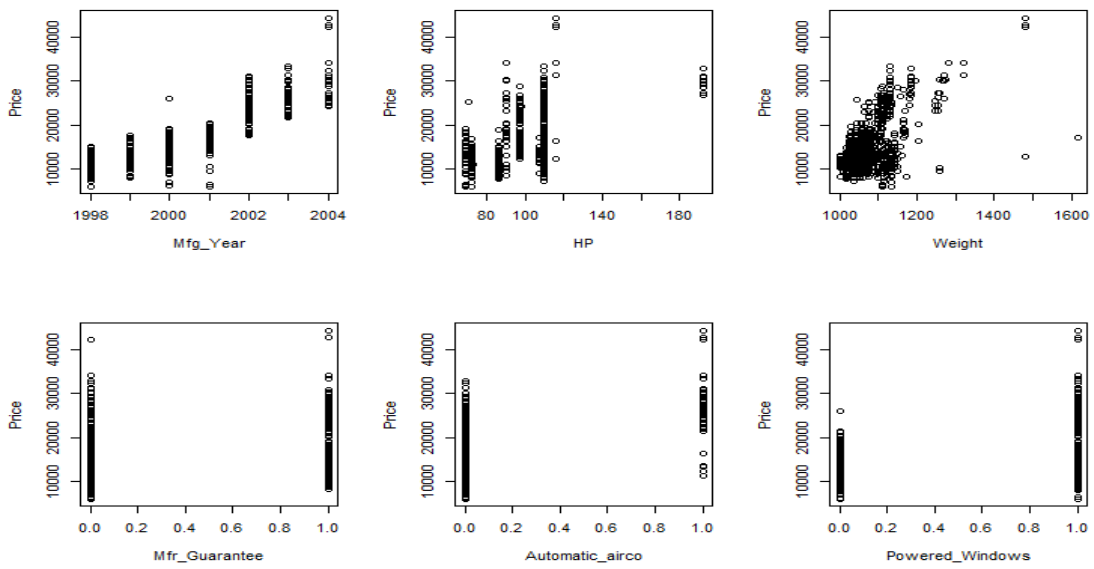


Figure 4: Scatterplot matrix (Figure 2) validates linearity assumption

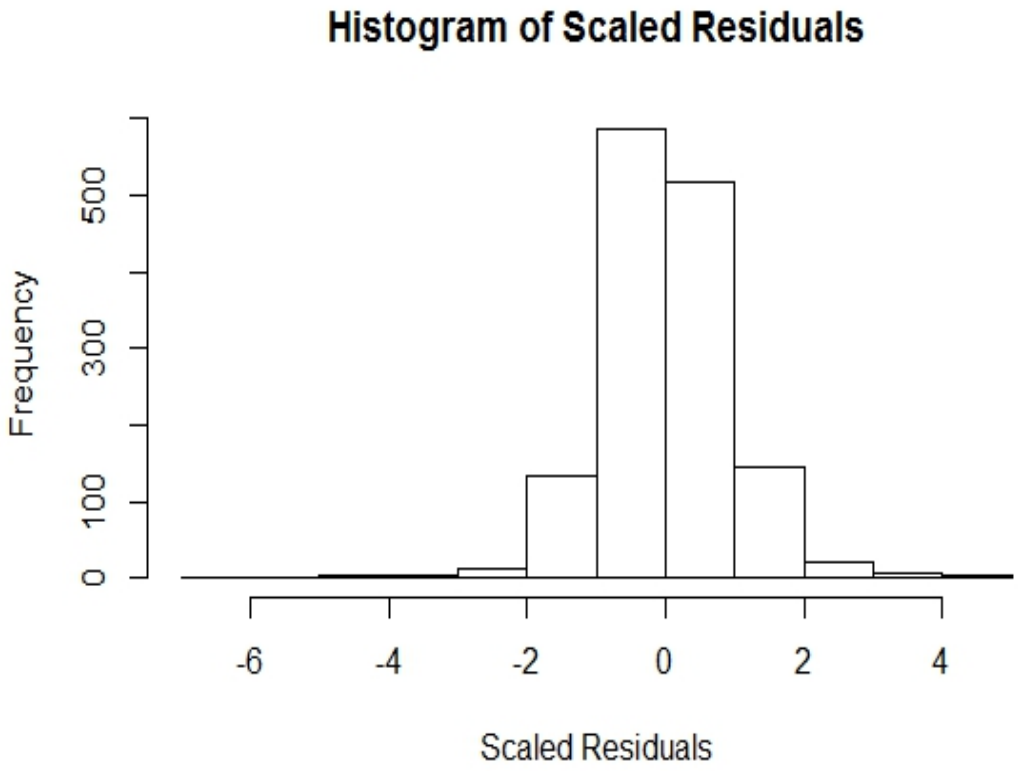


Figure 5: Histogram of residuals shows concern for normality assumption

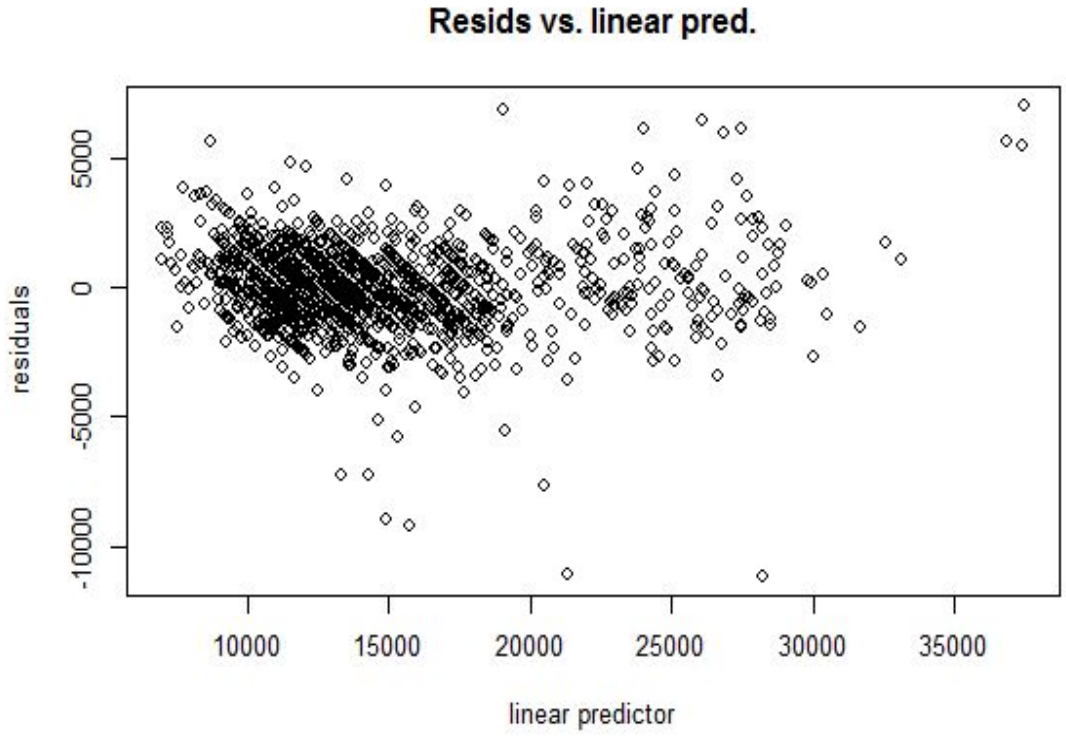


Figure 6: Plot of residuals vs fitted values validates equal variance assumption

Of particular interest is Confidence Interval coverage, which should be 95% but may have been effected by the lack of normality in the residuals.

Table 1: Prediction Diagnostics

	Estimate	Lower	Upper
Bias	-28.773	-112.147	54.602
MSE	2,341,443.000	2,227,397.000	2,455,489.000
RMSE	1,530.177	1,492.743	1,567.612
Coverage	0.967	0.966	0.969
Interval Width	6,045.905	6,044.992	6,046.818

Notice that although some of these results seem quite large, the Price variable was also large and quite variable, and this may have some impact upon the results. Some of these intervals are more tight than others which may reflect the lack of meeting of all the assumptions.

While our estimates are not exactly what we would expect, they are close, and from this we determine that our model predicts well on any given test set.

4 Results

This report adequately answers the questions posed in the case study, because we have used the data from the used car dealership in order to create a model that predicts the future sale price of any used car. The model is based upon natural splines and linear functions added together, yet this model seems to work will

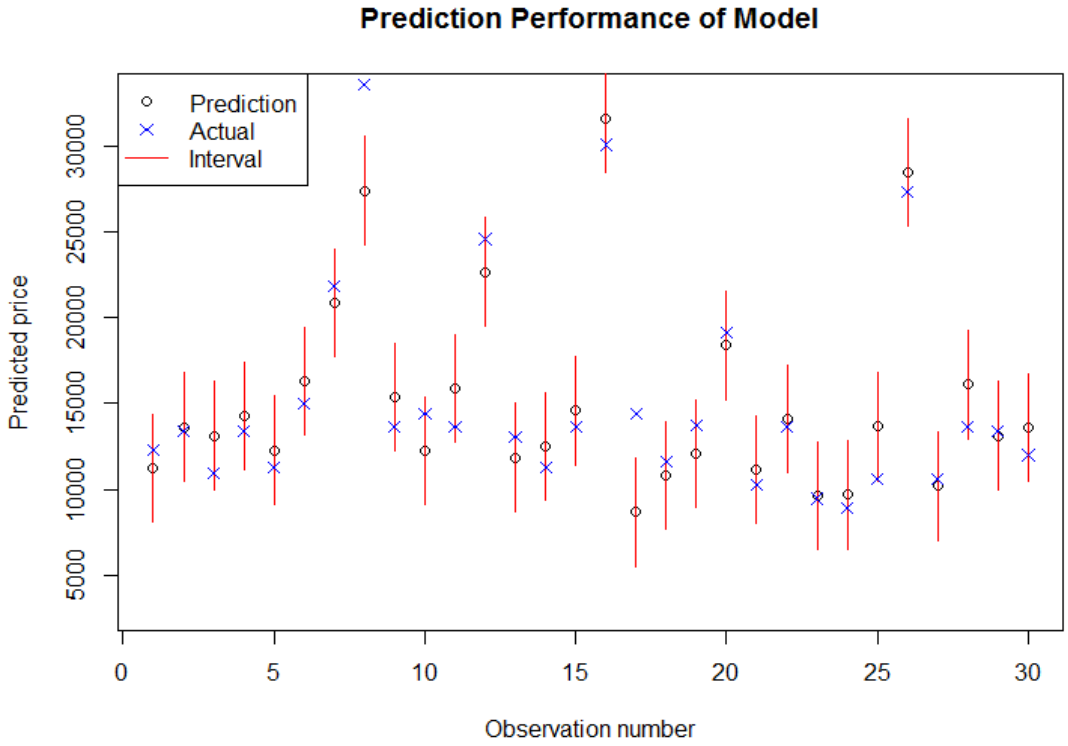


Figure 7: Assessment of prediction with n=30 observations in the test set. Note that Confidence Interval coverage is 28 out of 30, 93.3%, but due to variability in these simulations, this number is not as informative as the coverage calculated in Table 1.

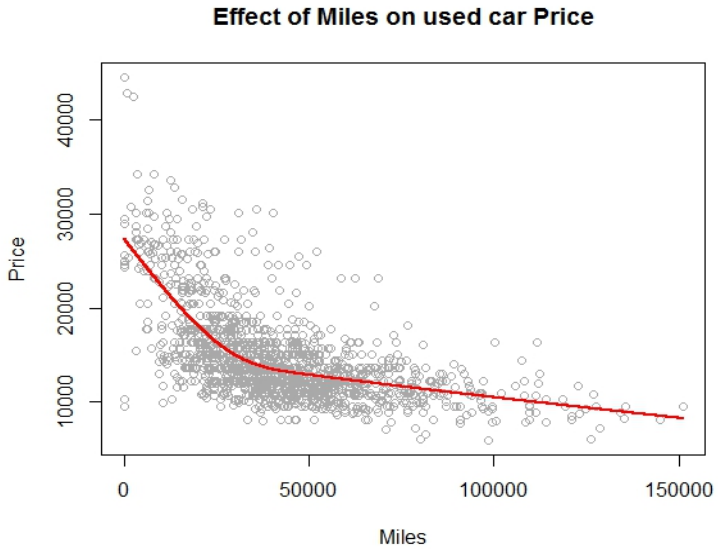
with a variety of combinations of test sets generated from the real data set.

Since we used a natural spline for the Miles variable as it appeared to have a nonlinear relationship with the Price response variable from our scatterplot as shown earlier, it is common to interpret a graph representing the relationship with Price and Miles and the effect of the natural spline.

The effect of miles on price is decreasing.

Table 2: Parameter Estimates

	bhat	lower	upper
(Intercept)	-3,728,966.000	-3,893,027.000	-3,564,906.000
Mfg_Year	1,861.402	1,779.086	1,943.718
HP	26.094	19.833	32.356
Weight	19.463	17.482	21.443
Mfr_Guarantee	415.363	239.692	591.033
Automatic_airco	3,317.640	2,885.639	3,749.640
Powered_Windows	629.884	449.155	810.614



As shown in the graph of Price vs. Miles with the natural spline overlaid, it appears that the as the number of miles increases, the price decreases. In fact, this is what we would expect to happen, because usually cars who have been driven more miles are likely to be either well used or have more problems or both. In such cases, customers are less likely to buy the car and so we see a decrease in price. Notice also from the plot of Miles and Price with the natural spline, that the Price drops quickly soon after Miles increase from zero, but around 50000 miles, the price decreases at a slower rate than before. This seems to indicate that when a car has been driven a few miles, the amount of the drop in price is significant, but after reaching a large threshold of Miles, the decrease is less significant.

The other explanatory variables included in the model: Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows are linear relationships with Price. Therefore, we can interpret the overall intercept from the Generalized Additive Model as well as each one of the β coefficients.

The parameter estimates are shown in 2. We found the estimates for the coefficients as well as 95 % confidence intervals for each of the coefficients with the t-distribution and standard errors from the GAM model. These confidence intervals show our uncertainty. If the process of generating intervals was repeated many times, on average 95 percent of the intervals would contain the true value for the intercept and each of the listed slopes.

However, it must be noted that these estimates are quite wide, because the Normality assumption of the model was not completely met. There were a few outliers in the data set in terms of Horsepower and Weight that are likely affecting the Normality condition of the data set. Further investigation should be done with the dealership to determine the cause of these outliers.

The intercept is quite large, because it includes the effect of the year variable where the year is measured

in the 2000s and so the intercept has to scale up for that. This means that if a car had zero of the explanatory variables, then the price would be $-\$3,728,966.000$. The interpretation of the intercept is not quite reasonable in the context of this problem. Instead, we focus on the predictions that this model is able to make, because that is the purpose of the problem.

We will interpret one of the β coefficients for an example with the others following similar interpretations.

For example, as the Manufacturing Year increases by one, the price will increase on average by $1,861.402$. Also, we are 95% confident that the true slope parameter for the Manufacturing Year lies between the interval of $(1,779.086,\ 1,943.718)$.

The other variables and their respective estimates for their contribution to the slope with their confidence intervals on the slope found from the standard errors of the slope coefficients on a t-distribution have similar interpretations as previously interpreted for Manufacturing Year.

Simply stated, the main points of the results are that we found an overall additive model for predicting the used car sale price. In this model, we included a nonlinear relationship of Miles and found that as the Miles increase the Price decreases. The other slope values for the variables of Manufacturing Year, Weight, HP, Manufacture Guarantee, Automatic Air-conditioning, and Powered Windows showed positive values, because the overall model included a very large negative intercept. The model also includes some random error to account for the uncertainty in the data and statistical method as well as to allow for the possibility of prediction by not creating a perfect fit of the data.

We found parameters which model the linear relationship of these variables - their slope coefficients and overall intercept with associated standard errors and confidence intervals. We also found a method to account for the non-linearity in the Miles variable and included this in the overall additive linear model. Finally, we found that despite some minor violations in the assumptions of the general additive linear model, the model performs well in predictions of used car selling price.

5 Conclusion

In summary, the goals of the analysis were met, because we created a model that is useful for predicting the selling price of used cars given some information about them, such as their Manufacturing Year, Weight, Miles, Horsepower, Manufacture , Automatic Air conditioning, and Powered Windows. It appears that those variables are the most significant of all the possible variables and characteristics about the used cars that are available in the data set about the used cars. It would be interesting to see if the model and associated explanatory variables would change given more current data. The newest car in this data set was made in 2004, which was 12 years ago. Perhaps people would be more interested in a Board Computer with the recent increases in technology or a USB connection with the spread of smartphones. Or it may be that given another more current data set that these variables would still be the most important in determining selling price.

There are a few relevant shortcomings with this model and approach used. First, we were surprised that after running our model with Cross-validation approaches and Best-subset selection and even accounting for non-linearity in some variables by using a natural spline or using some variables as factors instead of continuous that the Normality conditions of the model were still not completely met. We believe this was due to the presence of some influential observations in the Horsepower and Weight variables.

However, in future work, we could consider adding interactions between the variables into the model. This would need to be done carefully with best subset selection in an effort to not over- the model with linear dependencies. We found that as we would add more variables to the model as factors, that some of them would be linearly dependent with each other and have to be combined. It is difficult work for a statistician to combine all of these seemingly important variables and it would be more helpful to know which interactions should be investigated instead of needing to do it all by hand. Although we did not consider interactions in this case, they could be considered with a similar process of function selection or also with two-dimensional or splines.

For further research, investigation could be made into including more variables which exhibited an almost non-linear pattern into the model. Better methods for function selection could be developed so that the number of degrees of freedom for a natural spline could be tested in conjugation with the function selection or different functions of the same variable could be used in the best subset selection to know which truly is

the best function to use for each variable. We could also investigate how well natural splines do in comparison to other methods, such as wavelets, smoothers, or local regression. These methods all seem to have different advantages and it would be interesting to learn more about their similarities.