Project2

Connor Demorest

11/02/2020

Introduction

The data for this project is 'MNUsedCars2019', which is a snapshot of cars listed on http://findcars.com in March 2019. I originally created this dataset for a professor in undergrad for use in his regression analysis class.

Findcars.com is a regional car listing website where dealerships will list their cars they have for sale in the IA-WI-MN area. These data are 711 used sedans for sale by dealerships in March 2019. The outcome variable will be 'Price', the price of the car, with several possible explanatory variables including miles on the car, how many years old the car is, the make of the car, size of the car, and the color. The goal of this project is to model the price of used cars in the IA-WI-MN area listed online by dealerships.

Data Overview

The figures below are some exploratory data analysis to determine how candidate models should be fit and what transformations to continuous variables might be useful to make them closer to normally distributed.

Table 1: Summary statistics of used cars

	Mean	Std.Dev	Min	Q1	Median	Q3	Max
Price	11004.1	6205.9	795.0	5900.0	9995	15500.0	29995
Miles	89.5	61.4	2.5	36.9	76	133.2	318
AgeYearsOld	7.6	4.8	1.0	4.0	6	11.0	22

Table 2: Frequency of used car colors

Color	Freq	% Total
Other	102	14.3
Black	127	17.9
Blue	68	9.6
Gray	82	11.5
Red	50	7.0
Silver	159	22.4
White	84	11.8
Not Available	39	5.5
Total	711	100.0

Table 3: Frequency of used car makes

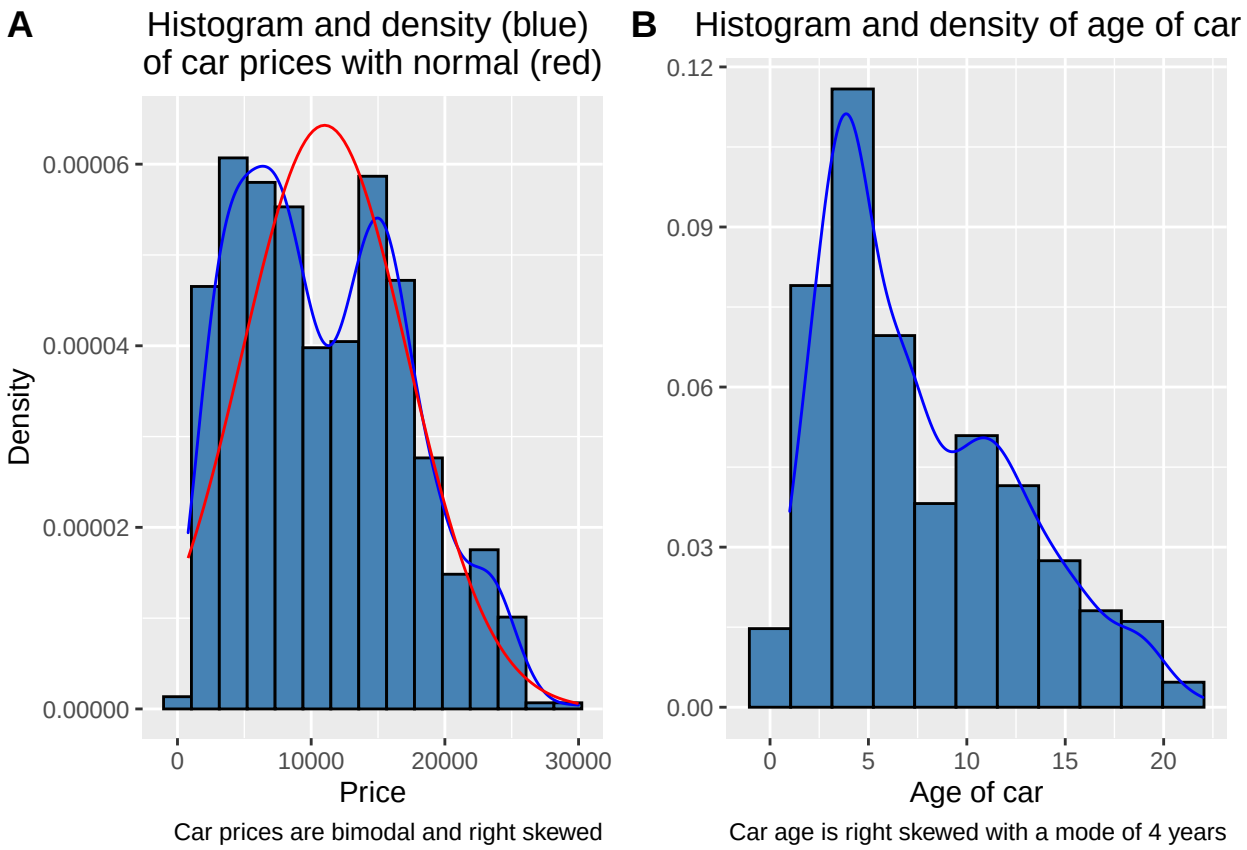
Make	Freq	% Total
Chevrolet	433	60.9
Honda	136	19.1
Toyota	142	20.0
Total	711	100.0

Table 4: Frequency of used car sizes

Size	Freq	% Total
More Space	300	42.2
Less Space	411	57.8
Total	711	100.0

Tables 1 through 4 show summary statistics of the variables in the data. Two things to note is that about 60% of the cars are Chevrolet, with the rest split evenly between Honda and Toyota, and that miles is measured in thousands of miles.

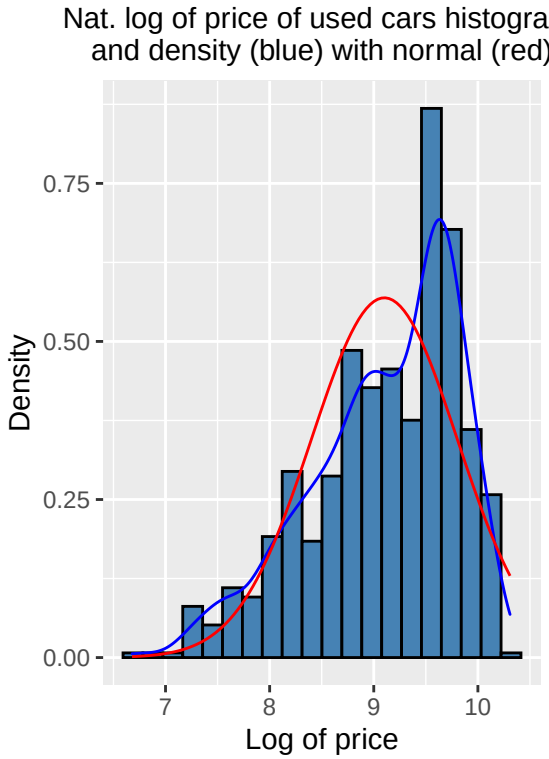
The outcome variable for this project is the car price in dollars. Figure A shows that the car prices are both bimodal and right-skewed. The car prices being right-skewed would be expected since the price has to be positive but there is no upper bound on prices. It's not clear why the prices are bimodal, but there are several potential explanations. There could be an overrepresentation of cars that were leased new for 3-4 years and then sold as used cars, and Figure B shows that the two most common ages of cars is 4 years followed by 3 years which would support that hypothesis. Another explanation would be the data collection process or sample itself is flawed, where some price points of cars would be more available than others for some different reason.



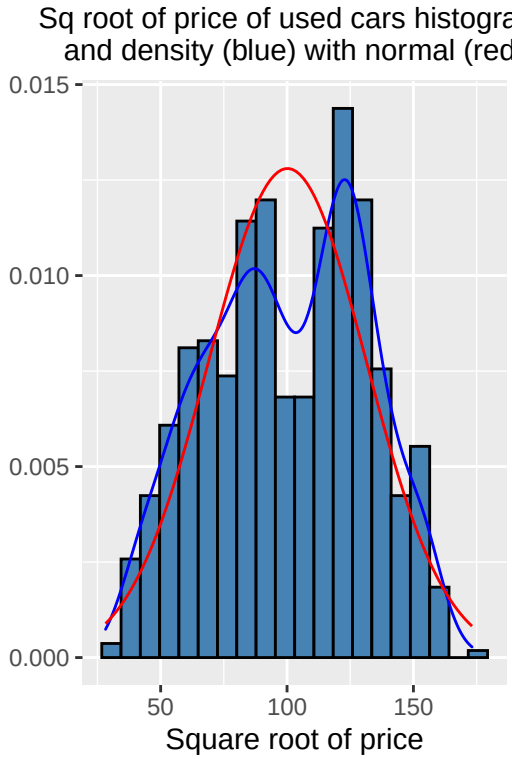
Potential predictor variables include number of miles on the car (in thousands), the make and model of the car, the age of the car (in years), and the color of the car. Each of those predictors was investigated for a bivariate relationship with the price of the car.

Figure C and D show a natural log transformation and square root transformation respectively to correct the skewness of used car prices. It seems like a square root transformation would be in order to make the prices of the used cars more normally distributed. It's not clear how important it is for the response variable to be normally distributed anyways, but normal is probably better than not-normal. The residuals being normally distributed is more important than the response being normally distributed.

C



D



It's pretty clear that used car prices explained by number of miles is not a linear relationship, there is curvature that isn't explained by the model shown in Figure E. One of the model assumptions of linear regression is that the relationship between the outcome and the predictors is linear, which is violated by modeling price with miles since there is curvature in the data in the scatterplot. This model would underpredict most cars with almost no miles and lots of miles and overpredict cars with a medium number of miles, and would predict cars with more than 220,000 miles to be worth less than 0 dollars, which is not likely since most car dealerships don't pay you to take their cars.

Figure F shows a modeling scheme using the square root transformation of price reduces the amount of curvature in the scatterplot, which leads to more accurate estimates of the price.

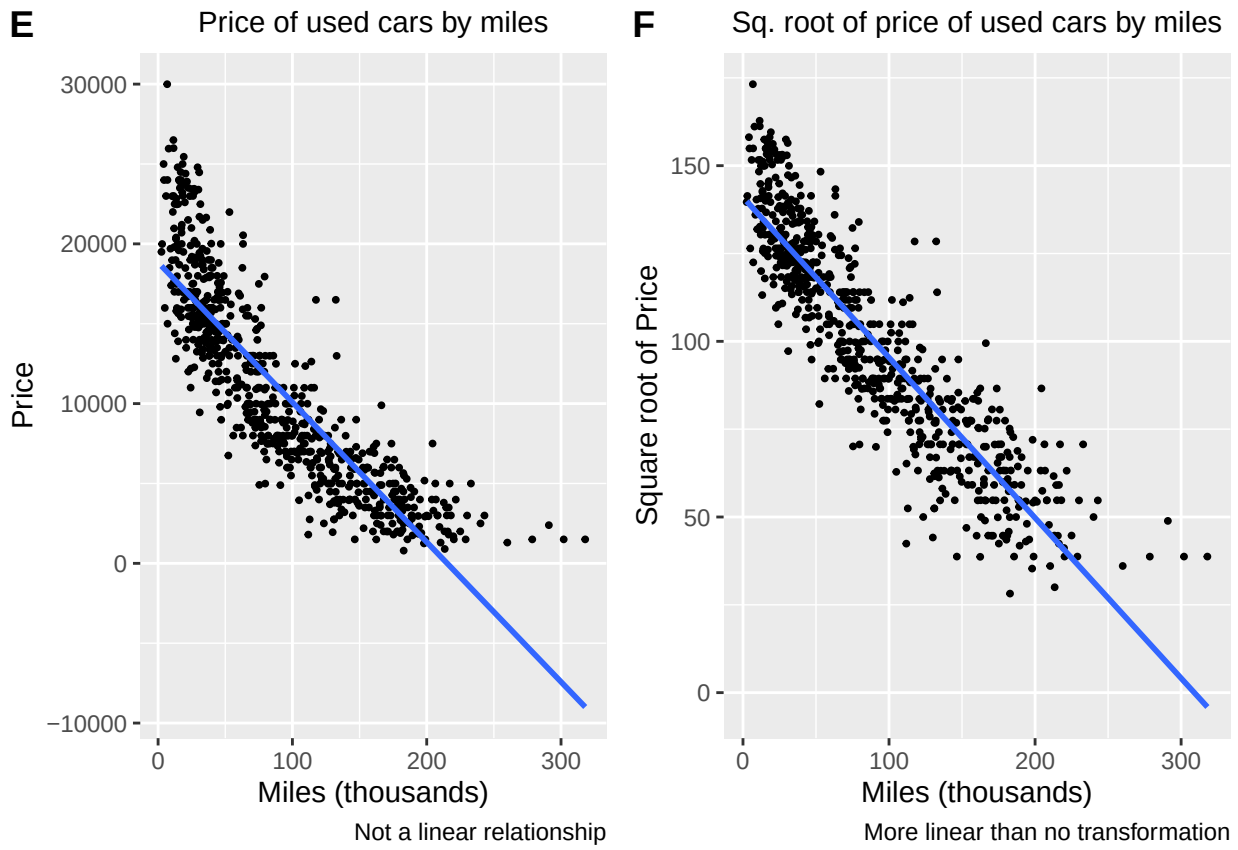
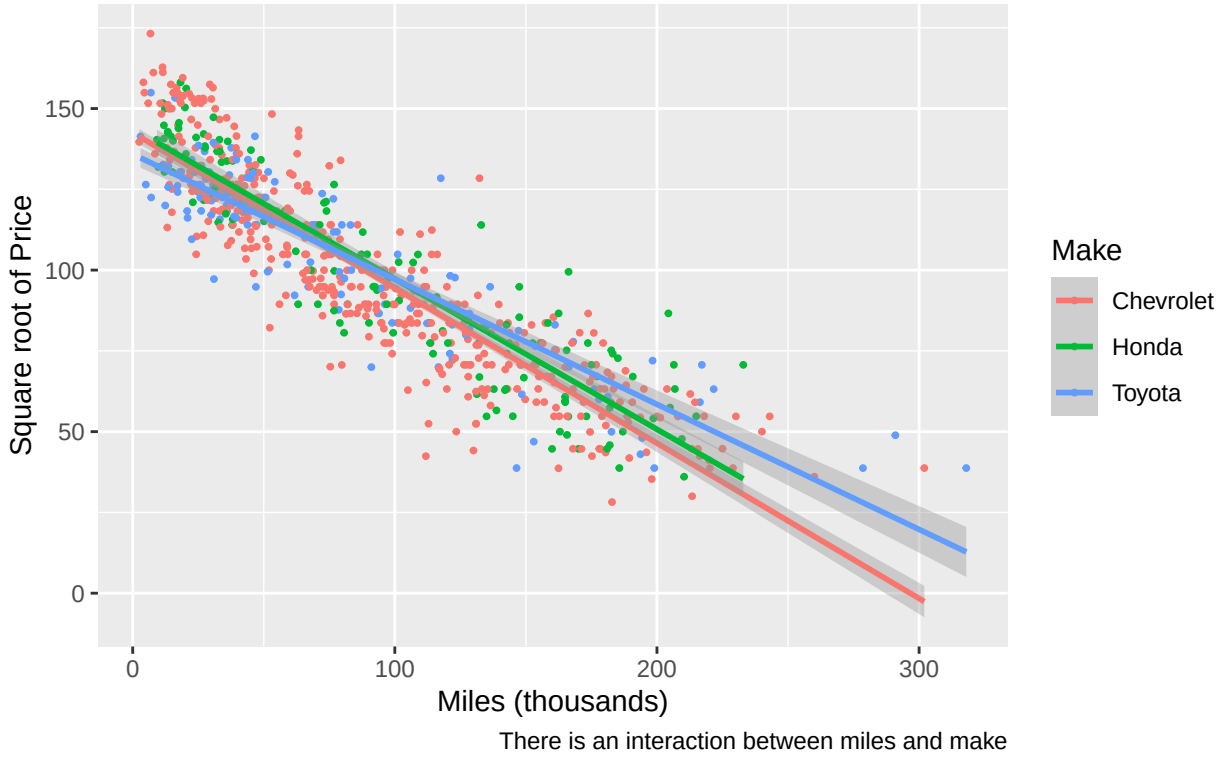


Figure G is an ANCOVA plot of the price, miles, and make of the used cars. A final model might include the amount of space in the car and the age of the car, but this is basically the gist of the model. This plot shows the interaction between the make of the car and the number of miles driven on them, which could be interpreted as how much better do some makes hold their value compared to the others at higher mileages.

There is clearly an interaction between the make and the square root of miles between Toyota and the other two makes of cars, indicated by different slopes between best fit lines with no overlap of the shaded standard error regions of the best fit lines. Toyota cars appear to hold their value better than the Chevrolet cars at high mileage, even though they are around the same price at low mileages.

Figure G: ANCOVA plot of square root of price by miles driven colored by make of the car



Statistical Procedures

Modeling 2019 MN used car price was done by fitting several candidate models and choosing between them using the estimated σ parameter value and using the fitted vs residual plots to make sure the model is doing what it should.

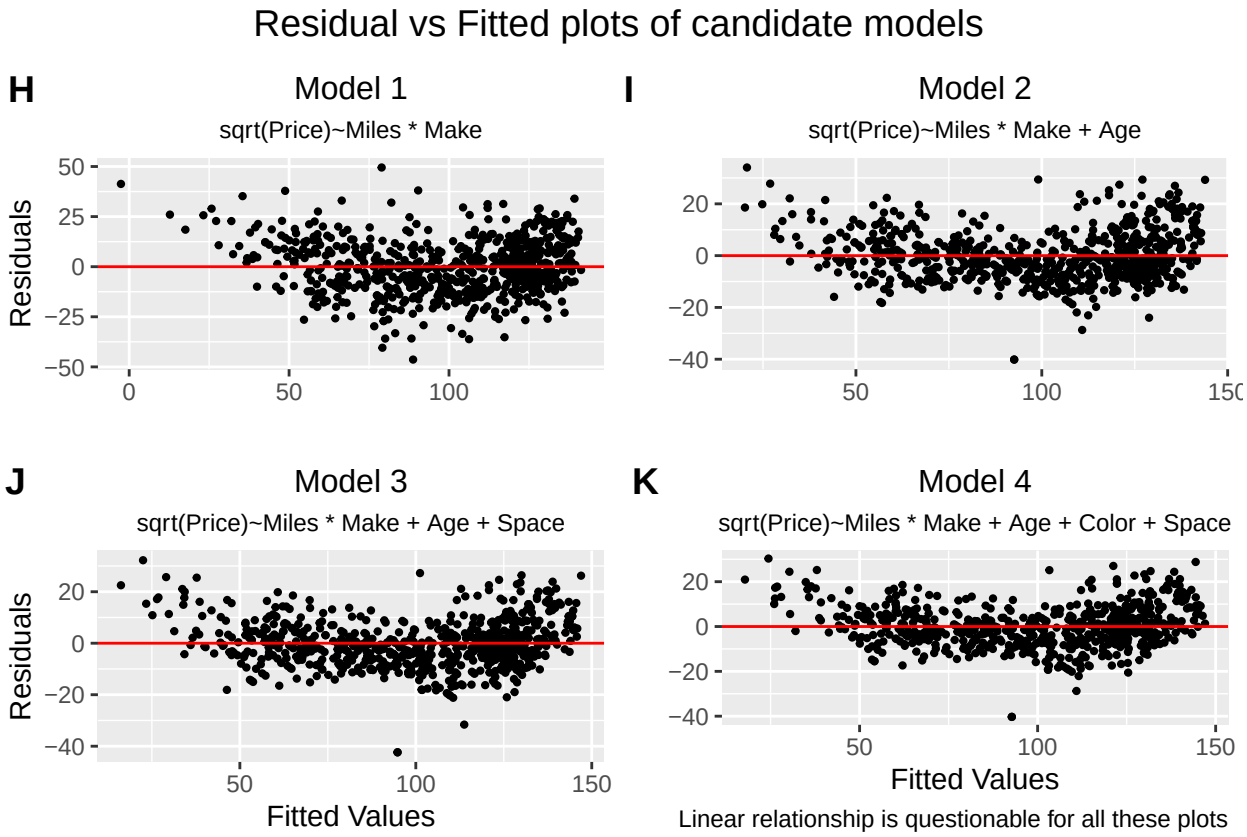


Table 5: Model diagnostics of candidate models

Model	Formula	Sigma	Sigma.SE
Model 1	Miles * Make	13.483	0.006
Model 2	Miles * Make + AgeYearsOld	9.723	0.004
Model 3	Miles * Make + AgeYearsOld + Space	9.258	0.004
Model 4	Miles * Make + AgeYearsOld + Color + Space	9.269	0.004

Table 6: Coefficients of Model 3 predictors

Predictor	Coefficient	Coefficient SE	CI Lower Bound	CI Upper Bound
Intercept	146.26	0.94	144.39	148.13
Miles	-0.28	0.01	-0.31	-0.26
Honda	0.27	1.71	-3.15	3.69
Toyota	-6.45	1.45	-9.35	-3.55
Age (years)	-3.38	0.13	-3.64	-3.12
Space (more room)	6.00	0.68	4.63	7.36
Miles:Honda	0.06	0.01	0.03	0.09
Miles:Toyota	0.11	0.01	0.08	0.14

The residual vs. fitted plots for all four candidate models are shown in plots H through K. Table 5 shows the estimated error values for the four candidate models. Based on both the residual vs. fitted plot and the estimated error values, Model 3 is the best model out of the candidates. It has 4 predictors, all of them except for the color of the car. The coefficients of the model is shown in table 6. Unfortunately, the problem with

using the square root transformation for the price is that there is no nice interpretation of the coefficients of the model predictors in the original scale of price. The coefficients can be compared to each other. For example, the car having more space is associated with the same change in price as a car with approximately 21,400 fewer miles, if everything else is the same.

The final model: $\sqrt{(Price)} = \beta_0 + \beta_1 * Miles + \beta_2 * Honda + \beta_3 * Toyota + \beta_4 * Age + \beta_5 * MoreSpace + \beta_6 * MilesifHonda + \beta_7 * MilesifToyota + error$

Results

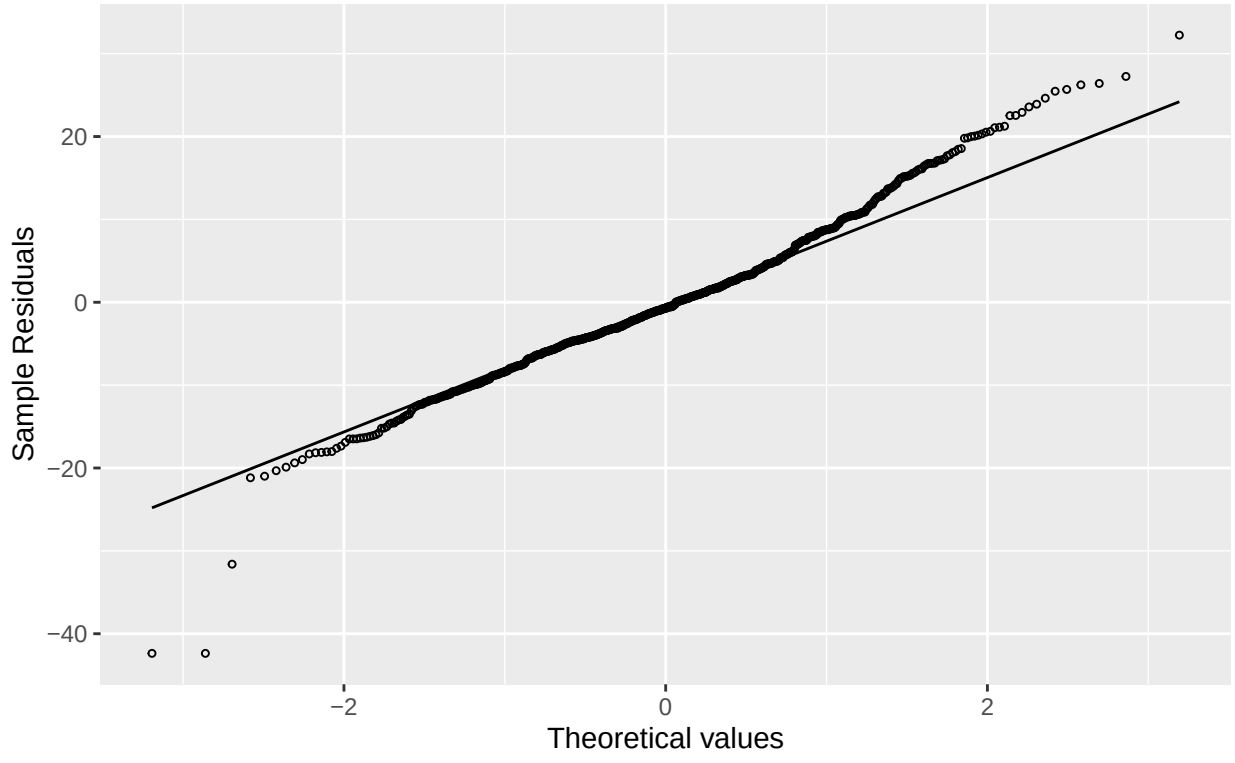
In the final model, the estimated standard deviation of the error for all candidate models is diplayed in table 5. Table 6 shows that the best model (Model 3) had a estimated error of 9.258 (95% interval: 9.250, 9.266). With that estimate of how much error there was in the model, there are several model assumptions.

Linearity: Linearity would be satisfied if the values are centered around zero and there is no curvature in that plot. Figure J was used to assess the linearity assumption of the model. There appears like there could be some curvature at the tail ends of the fitted values, for both high and low values of predicted used car prices, but for the middle part of the data, there is equal and random spread around the red line at 0.

Homoskedasticity: Figure J was used to assess if the residuals have constant variance. Homoskedasticity would be violated if there is a fan pattern in the residual vs predicted plot. There are no major amounts of change in the amount of variance in the residuals as the fitted values increase.

Normally distributed residuals: To assesss if the model has normally distributed residuals, a normal-quantile plot of the residuals should show the data following the trend line and do not stray too much. Figure L shows a normal-quantile plot of the data. The residuals tend to follow the theoretical values mostly but it tends to have a longer upper tail (i.e. right skewness) that would not follow a normal distribution exactly.

Figure L: Normal −Quantile plot of Model 3 residuals



No autocorrelation in the data: The assumption of not having autocorrelation in the data is a tricky

discussion. I touched on it earlier in the report, in the Data Overview section. The data were collected from an online website that aggregates the used car data for the dealerships that pay the site's owners to put it on there. It's hard to say if the car prices would be autocorrelated for sure one way or the other because the price of a certain local car might be affected by the prices of other local cars, especially if they are similar. I would argue that even if this effect exists, it should be relatively small in relation to the total amount of variation in the data due to unexplained error, and could be ignored for these purposes.

The fitted model with the estimated β values included:

$\widehat{\sqrt{Price}} = 146.26 - 0.28 * Miles - 0.27 * Honda - 6.45 * Toyota - 3.38 * Age + 6.00 * MoreSpace + 0.06 * MilesifHonda + 0.11 * MilesifToyota$

For a Chevy car with 0 miles, less space, and zero years old, the estimated price of a car would be \$21,400, which actually seems fairly accurate, especially considering that the data used was only used cars and did not include any new cars (a new 2020 Malibu is \$22,000 for a base model). https://www.chevrolet.com/cars/previous-year/malibu

The interaction between Toyota and Honda cars and miles indicates that even though Chevy cars might start out at a higher price than both Hondas and Toyotas, when keeping everything else constant, the prices of Honda and Toyota cars decrease less quickly than Chevy cars.

For example, on a Toyota car with 100,000 miles, less space, and 8 years old, the estimated price of that car is \$9172. To compare, a Honda with those same parameters would be estimated to be \$9400, and a Chevrolet would be \$8321. At 200,000 miles and 15 years old with less space, a Toyota would be estimated at \$3037, a Honda would be \$2630, and a Chevy would be \$1565.

My old car would be a super-outlier, if that's a word. It was a 1999 Buick Lesabre with 87,000 miles on it and I would guess "more space". I got it from my great-grandmother when she passed, she bought it and really only took it to church for a long time. It died on the move out to Bozeman, but the model would predict that it's worth \$3240 if I assume that it is basically a Chevrolet (both are General Motors made is what I figured). That's really close to what my parents were offered for it when they looked into selling it. That's pretty cool!

Discussion

Model Limitations: This model is not perfect, of course. It can make predictions for cars with very high mileage in the negative amounts of money, meaning that the dealership would pay you to take their car, which of course makes no sense. That's why in the residual vs predicted plot, the residuals of all the cars near predicted prices of \$0 have positive residuals, because the model is not great at handling cars with high mileage. For many cars in the middle part of the number of miles, this model is pretty good at estimating the price that would be fair for a IA-WI-MN used car with that many miles.

Scope of Inference: Hopefully, the findings from this model could be used to determine which cars might have a listed price under market value for the characteristics of the car. It's hard to tell what the scope of inference might be for this model, since the used car market is always changing. Ideally, this model could be used for used cars in any year in the IA-WI-MN area sold by dealerships.

Future work: I don't know if this is required, but I was thinking that a MARS model would be really interesting for these data, because it would fit a model for cars with low mileage, a different model for cars with medium mileage, and another for cars with high mileage, which would help with the problem that I pointed out in the limitations section.

Next time I would use log transformation of car prices or something else for homoskedasticity and linearity. I was interested in the interpretations of the coefficients of the predictors, but when the outcome is in the square root scale, a 1 unit change in a predictor when holding everything else constant doesn't have much of an interpretation. If I had kept the original scale, it would have an effect in real dollars, or if I used the natural log transformation, it would be a multiplicative increase in the price in dollars.

Acknowledgments

Thank you to Gabby Lemire, Andrew Abraham, and Therese Lupariello for giving me feedback on several parts of the project, especially the exploratory data analysis section.

Citations

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. Regression and Other Stories. Cambridge University Press.

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Hadley Wickham (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats

Dominic Comtois (2020). summarytools: Tools to Quickly and Neatly Summarize Data. R package version 0.9.6. https://CRAN.R-project.org/package=summarytools

Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 https://mc-stan.org/rstanarm.

Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J & Wolfe R. Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA. https://github.com/stan-dev/stancon_talks/

Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.0. https://CRAN.R-project.org/package=cowplot

Appendix

```
knitr::opts_chunk$set(echo = FALSE, fig.align = 'center', message = F, warning = F)
library(knitr)
library(tidyverse)
library(forcats)
library(summarytools)
library(rstanarm)
library(cowplot)


# Recode as factors
cars = read_csv(file = "MNUsedCars2019.csv") %>%
  mutate(Make = as.factor(Make),
         Model = as.factor(Model),
         Space = as.factor(Space),
         State = as.factor(State),
         # Make NA a level of the factor
         Color = as.factor(Color),
         # Scale Miles to thousands to increase effect size
```

```
        Miles = Miles / 1000
        )

# Combine less common colors
other_cols = c("Beige", "Bronze", "Brown", "Burgundy", "Champagne", "Gold", "Green", "Maroon", "Pearl",
cars$Color = fct_collapse(cars$Color, Other = other_cols)

# Make this... better
st_options(descr.stats = c("mean","sd","min","q1","med","q3","max"), descr.transpose = T, headings = F)
cont_vars = rbind(
descr(cars$Price, round.digits = 0),
descr(cars$Miles, round.digits = 0),
descr(cars$AgeYearsOld, round.digits = 1))

kable(x = cont_vars, digits = 1, caption = "Summary statistics of used cars")

make_cat = data.frame(Make = c("Chevrolet", "Honda", "Toyota","", "Total"),
                      freq(cars$Make, cumul = F, report.nas = F, totals = T))
color_cat = data.frame(Color = c("Other", "Black", "Blue", "Gray", "Red", "Silver", "White", "Not Availa
                       freq(cars$Color, cumul = F, report.nas = T, totals = T))
space_cat = data.frame(AmtSpace = c("More Space", "Less Space", "", "Total"),
                       freq(cars$Space, cumul = F, report.nas = F, totals = T))

kable(color_cat[,c(1,2,5)],
      digits = 1,
      caption = "Frequency of used car colors",
      row.names = F,
      col.names = c("Color", "Freq", "% Total"))
kable(make_cat[c(1:3, 5), c(1, 2, 5)],
      digits = 1,
      caption = "Frequency of used car makes",
      row.names = F,
      col.names = c("Make", "Freq", "% Total"))
kable(space_cat[c(1,2,4), c(1, 2, 5)],
      digits = 1,
      caption = "Frequency of used car sizes",
      row.names = F,
      col.names = c("Size", "Freq", "% Total"))


options("scipen" = 10)
p = ggplot(data = cars, aes(y = Price))


f1 = ggplot(data = cars) +
  geom_histogram(aes(x = Price, y = ..density..), fill = "steelblue", col = "black", bins = 15) +
  geom_density(aes(x = Price, y = ..density..), col = "blue") +
  stat_function(fun = dnorm,
                args = with(cars, c(mean = mean(Price), sd = sd(Price))), inherit.aes = F, col = "red")
  labs(x = "Price",
       y = "Density",
       title = "Histogram and density (blue) \nof car prices with normal (red)",
       caption = "Car prices are bimodal and right skewed") +
```

```
  theme(plot.title = element_text(hjust = 0.5))

f2 = ggplot(data = cars, mapping = aes(x = AgeYearsOld, y = ..density..)) +
  geom_histogram(bins = 11, fill = "steelblue", col = "black") +
  geom_density(col = "blue") +
  labs(title = "Histogram and density of age of cars",
       x = "Age of car",
       y = "",
       caption = "Car age is right skewed with a mode of 4 years") +
  theme(plot.title = element_text(hjust = 0.5))

cowplot::plot_grid(f1, f2, labels = c("A", "B"))


3 = ggplot(data = cars, mapping = aes(x = log(Price), y = ..density..)) +
  geom_histogram(col = "black", fill = "steelblue", bins = 20) +
  geom_density(col = "blue") +
  stat_function(fun = dnorm,
                args = with(cars, c(mean = mean(log(Price)), sd = sd(log(Price)))),
                inherit.aes = F, col = "red") +
  labs(title = "Nat. log of price of used cars histogram \nand density (blue) with normal (red)",
       x = "Log of price",
       y = "Density") +
  theme(plot.title = element_text(hjust = 0.5, size = 11))

f4 = ggplot(data = cars, mapping = aes(x = sqrt(Price), y = ..density..)) +
  geom_histogram(col = "black", fill = "steelblue", bins = 20) +
  geom_density(col = "blue") +
  stat_function(fun = dnorm,
                args = with(cars, c(mean = mean(sqrt(Price)), sd = sd(sqrt(Price)))),
                inherit.aes = F, col = "red") +
  labs(title = "Sq root of price of used cars histogram \nand density (blue) with normal (red)",
       x = "Square root of price",
       y = "") +
  theme(plot.title = element_text(hjust = 0.5, size = 11))

plot_grid(f3, f4, labels = c("C", "D"), scale = 0.9)


f5 = p + geom_point(aes(x = Miles), size = 0.7) +
  geom_smooth(aes(x = Miles), method = "lm", se = F, formula = y~x) +
  labs(x = "Miles (thousands)",
       y = "Price",
       title = "Price of used cars by miles",
       caption = "Not a linear relationship") +
  theme(plot.title = element_text(hjust = 0.5, size = 11))

f6 = ggplot(data = cars, mapping = aes(x = Miles, y = sqrt(Price))) +
  geom_point(size = 0.7) +
  geom_smooth(method = "lm", se = F, formula = y~x) +
  labs(title = "Sq. root of price of used cars by miles",
       x = "Miles (thousands)",
       y = "Square root of Price",
```

```
       caption = "More linear than no transformation") +
  theme(plot.title = element_text(hjust = 0.5, size = 11))

cowplot::plot_grid(f5, f6, labels = c("E", "F"))


p + geom_point(aes(x = Miles, y = sqrt(Price), col = Make), size = 0.7) +
  geom_smooth(aes(x = Miles, y = sqrt(Price), col = Make), method = "lm", formula = y~x) +
  labs(title = "Figure G: ANCOVA plot of square root of price by miles driven \ncolored by make of the
       x = "Miles (thousands)",
       y = "Square root of Price",
       caption = "There is an interaction between miles and make") +
  theme(plot.title = element_text(hjust = 0.5))
# There appears to be an interaction effect in the ANCOVA plot


set.seed(10172020)
m1stan = stan_glm(sqrt(Price) ~ Miles * Make , data = cars, refresh = 0)
f8 = ggplot(mapping = aes(x = m1stan$fitted.values, y = m1stan$residuals)) +
  geom_point(size = 0.8) +
  geom_abline(intercept = 0, slope = 0, col = "red") +
  labs(title = "Model 1",
       subtitle = "sqrt(Price)~Miles * Make",
       x = "",
       y = "Residuals") +
  theme(plot.title = element_text(hjust = 0.5, size = 11.5),
        plot.subtitle = element_text(hjust = 0.5, size = 9))

m2stan = stan_glm(sqrt(Price) ~ Miles * Make + AgeYearsOld, data = cars, refresh = 0)
f9 = ggplot(mapping = aes(x = m2stan$fitted.values, y = m2stan$residuals)) +
  geom_point(size = 0.8) +
  geom_abline(intercept = 0, slope = 0, col = "red") +
  labs(title = "Model 2",
       subtitle = "sqrt(Price)~Miles * Make + Age",
       x= "",
       y = "") +
  theme(plot.title = element_text(hjust = 0.5, size = 11.5),
        plot.subtitle = element_text(hjust = 0.5, size = 9))

m3stan = stan_glm(sqrt(Price) ~ Miles * Make + AgeYearsOld + Space, data = cars, refresh = 0)
f10 = ggplot(mapping = aes(x = m3stan$fitted.values, y = m3stan$residuals)) +
  geom_point(size = 0.8) +
  geom_abline(intercept = 0, slope = 0, col = "red") +
  labs(title = "Model 3",
       subtitle = "sqrt(Price)~Miles * Make + Age + Space",
       x = "Fitted Values",
       y = "Residuals") +
  theme(plot.title = element_text(hjust = 0.5, size = 11.5),
        plot.subtitle = element_text(hjust = 0.5, size = 9))

m4stan = stan_glm(sqrt(Price) ~ Miles * Make + AgeYearsOld + Color + Space, data = cars, refresh = 0)
f11 = ggplot(mapping = aes(x = m4stan$fitted.values, y = m4stan$residuals)) +
  geom_point(size = 0.8) +
```

```
  geom_abline(intercept = 0, slope = 0, col = "red") +
  labs(title = "Model 4",
       subtitle = "sqrt(Price)~Miles * Make + Age + Color + Space",
       x = "Fitted Values",
       y = "",
       caption = "Linear relationship is questionable for all these plots") +
  theme(plot.title = element_text(hjust = 0.5, size = 11.5),
        plot.subtitle = element_text(hjust = 0.5, size = 9))
# Overfitting?
# Could remove Color, none of the colors are practically significant. At most a 5% change in price, plu
# Investigate super outlier at fitted value = 8.9, actual value = 7.9?

# Fitted vs Resid plots
plots = plot_grid(f8,f9,f10,f11, labels = c("H","I","J","K"), nrow = 2)
title = ggdraw() + draw_label("Residual vs Fitted plots of candidate models")
plot_grid(title, plots, rel_heights = c(0.1,1), ncol = 1)

# Table of sigma values
mod_diagnostics = data.frame(Model = c(paste("Model", 1:4)),
                Formula = c(as.character(m1stan$formula)[3], as.character(m2stan$formula)[3], as.charac
                Sigma = c(summary(m1stan)[7,1], summary(m2stan)[8,1], summary(m3stan)[9,1], summary(m4s
                `Sigma SE` = c(summary(m1stan)[7,2], summary(m2stan)[8,2], summary(m3stan)[9,2], summar
kable(x = mod_diagnostics, digits = 3, caption = "Model diagnostics of candidate models")

# Table of best model coef
df = data.frame(Rownames = c("Intercept", "Miles", "Honda", "Toyota", "Age (years)", "Space (more room)
                Coefficient = m3stan$coefficients,
                Coefficient.SE = m3stan$ses,
                `Lower Bound` = m3stan$coefficients - 2*m3stan$ses,
                `Upper Bound` = m3stan$coefficients + 2*m3stan$ses)
kable(df, digits = 2,
      caption = "Coefficients of Model 3 predictors",
      col.names = c("Predictor", "Coefficient", "Coefficient SE", "CI Lower Bound", "CI Upper Bound"),
      row.names = F)


ggplot(mapping = aes(sample = m3stan$residuals)) + stat_qq(size = 0.9, shape = 1) + stat_qq_line() +
  labs(title = "Figure L: Normal -Quantile plot of Model 3 residuals",
       x = "Theoretical values",
       y = "Sample Residuals",
       caption = "Again, at higher car price the model starts to fail") +
  theme(plot.title = element_text(hjust = 0.5))
```