MATH 424: Multiple Linear Regression project

Nathan Oliver

Collaboration rules:

You may consult with up to two classmates for help with this project, but use your own data (must have different make/model/zip codes). Please identify who you collaborate with here:

Read this document before you submit it to ensure there is not a ton of extra output that does not contribute to the analysis or communication. Also, I recommend using the spell-checker in RStudio (Edit -> Check Spelling). Note that you will need to closely follow the instructions on the Canvas assignment page to complete this project successfully.

```
# clean-up R environment
rm(list = ls())

# Load Packages
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
```

```
##
## Attaching package: 'mosaic'
```

```
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
```

```
## The following object is masked from 'package:Matrix':
##
##     mean
```

```
## The following object is masked from 'package:ggplot2':
##
##     stat
```

```
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
```

```
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```r
library(ggformula)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x mosaic::count() masks dplyr::count()
## x purrr::cross()  masks mosaic::cross()
## x mosaic::do()    masks dplyr::do()
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x mosaic::stat()  masks ggplot2::stat()
## x mosaic::tally() masks dplyr::tally()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following objects are masked from 'package:mosaic':
##
##     deltaMethod, logit
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(tinytex)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
knitr::opts_chunk$set(echo = F)  #make it be a default to not include the code, only the output, when y

# suggestion for sourcing and wrangling data:
```

Introduction

In this project, I will be analyzing the price and mileage differences of the Nissan Maxima in Seattle (zip: 98101) and Chicago (zip: 60007). This vehicle was primarily chosen because I like Nissan and the Nissan

Maxima is a sedan that I've always found desirable. I chose Seattle because that's my favorite city, and I chose Chicago because I find it to be extremely different both culturally and economically. This data has been sourced from http://myslu.stlawu.edu/~clee/dataset/autotrader/ , and the data requests were limited to 300 entries. After acquiring the data sets, they were merged together into one large data set with any vehicles containing zero values removed. Additionally, a column was added that computed the age of each individual vehicle.

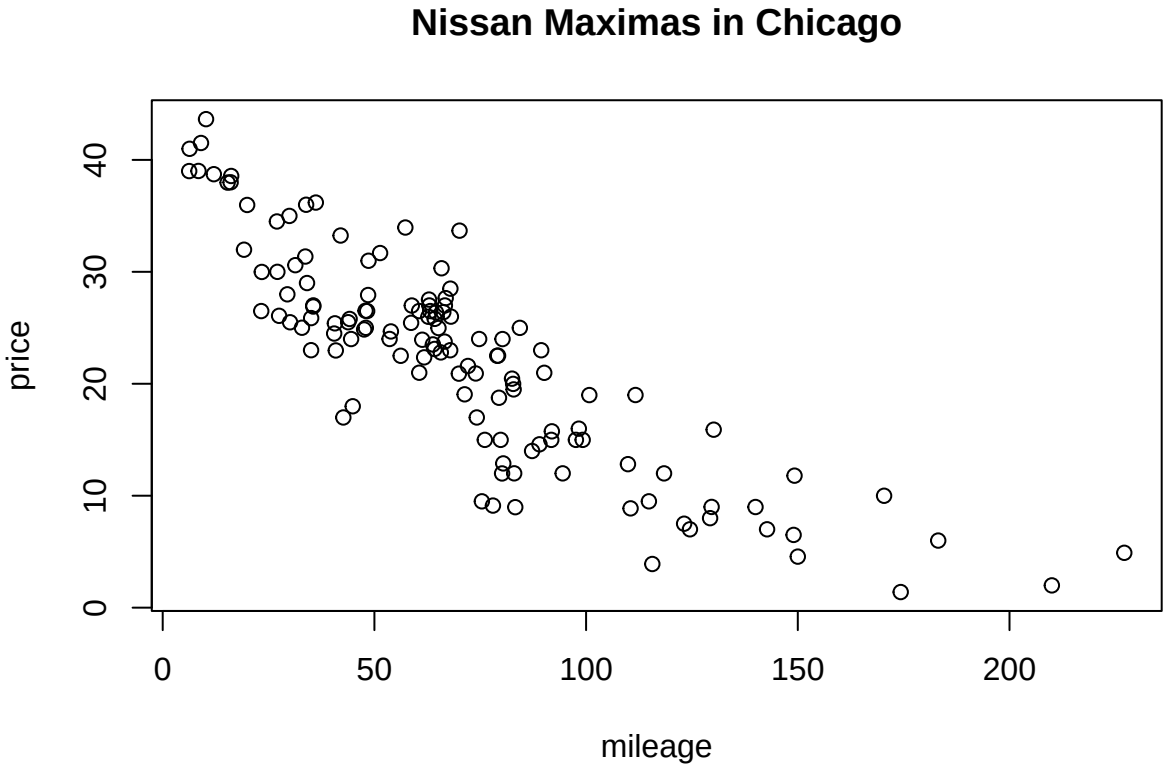
I think that the Nissan Maxima will be more expensive on average in Seattle. Seattle is a more expensive place to live, and is growing much faster than Chicago (which I believe is actually getting smaller these days). My assumption is that there is a much larger demand in Seattle over Chicago, resulting in higher prices.

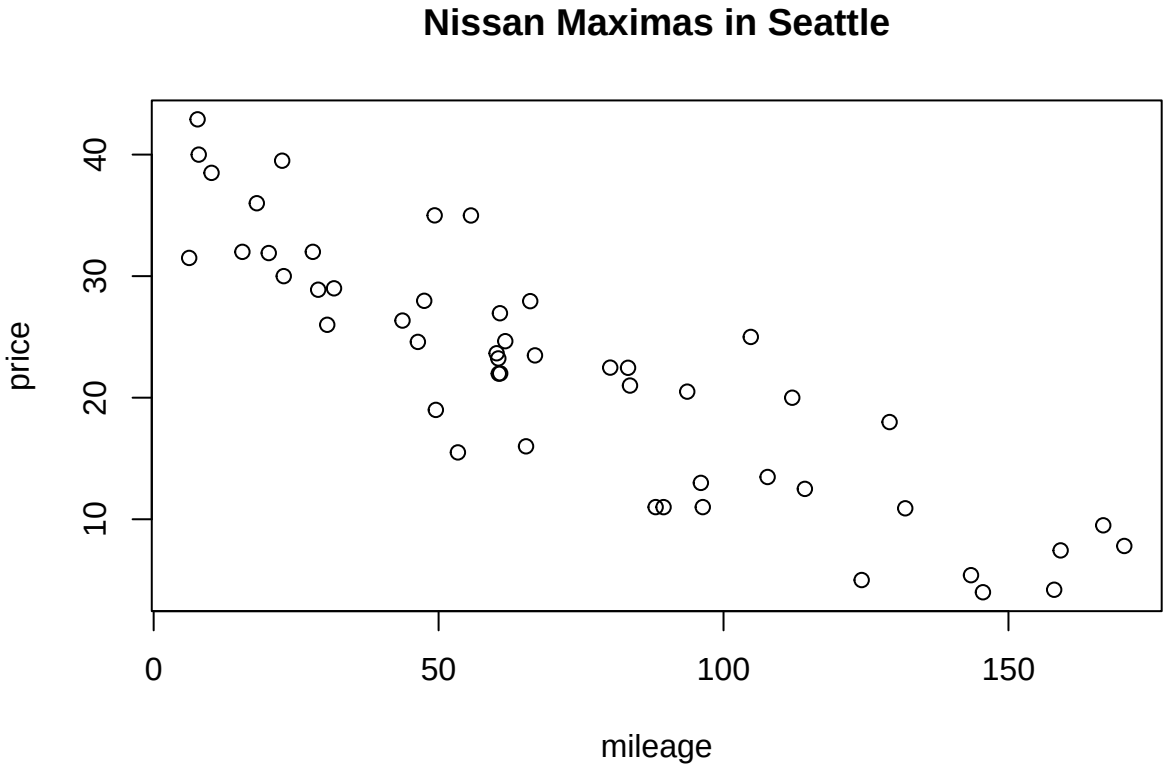
Research question 1

Assuming a linear relationship between price and mileage, is there a difference in price between the locations?

Exploratory data analysis

Figure(s):





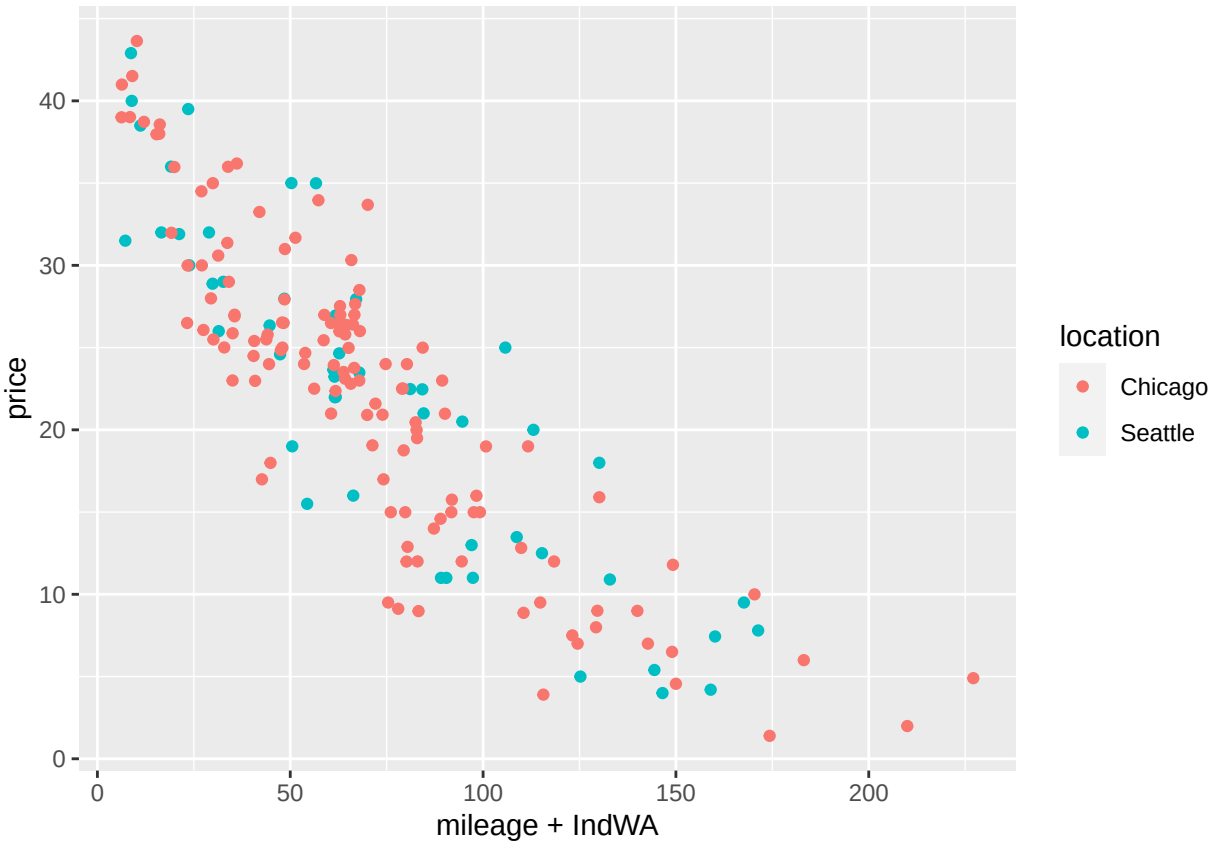
EDA TABLE 1

	sample size	mean price	sd of price	mean mileage	sd of mileage
Chicago	128	22.35	09.38	70.46	41.52
Seattle	50	22.14	10.30	72.72	46.00

Comments: Some of my first impressions of this data make me rather suspicious of any major differences. Both spreads of data look like they have almost the exact same intercept, and maybe a marginally different slope of regression. The mean prices between areas is almost identical. The standard deviation of price and mean of mileage are also remarkably similar. The only major difference of note is the standard deviation of mileage between areas. Even though the standard deviation of mileage of Seattle is a bit larger than Chicago, there is also notably less data. With more data from Seattle, we might see more of an impact from CLT that could result in a smaller standard deviation.

Model fitting

MODEL SUMMARY TABLE 1: (same slope different intercepts)

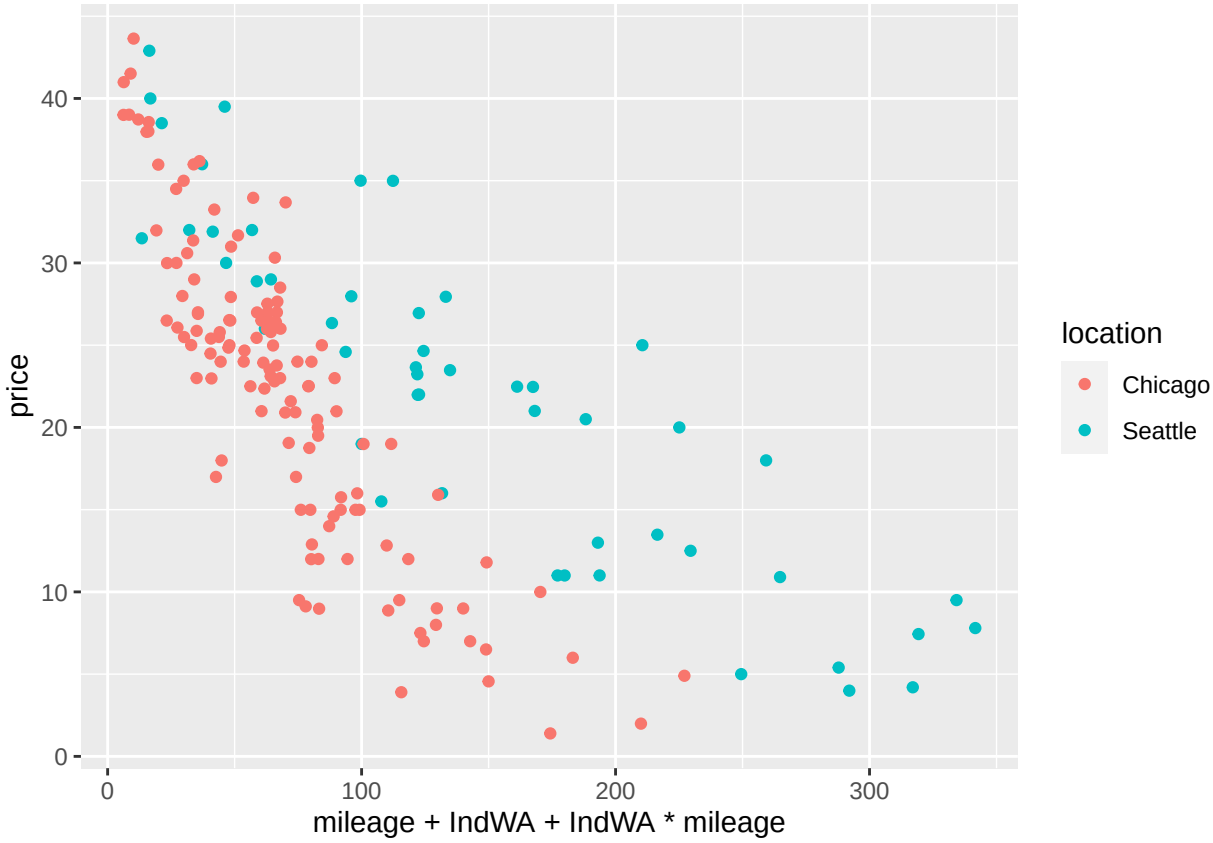



```
##
## Call:
## lm(formula = price ~ mileage + IndWA, data = true_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8853  -2.7524  -0.4458   3.6319  13.1692
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.116857   0.733315  49.252   <2e-16 ***
## mileage     -0.195441   0.008476 -23.059   <2e-16 ***
## IndWA        0.237331   0.803187   0.295    0.768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.815 on 175 degrees of freedom
## Multiple R-squared:  0.7524, Adjusted R-squared:  0.7496
## F-statistic: 265.9 on 2 and 175 DF,  p-value: < 2.2e-16
```

	estimate	test-statistic	p-value
intercept	36.12	49.25	<2e-16
mileage	-0.19	-23.059	<2e-16
location	0.237331	0.295	0.768

MODEL SUMMARY TABLE 2: (different slopes and different intercepts)

	estimate	test-statistic	p-value
intercept	36.12	49.25	<2e-16
mileage	-0.19	-23.059	<2e-16
location	0.237331	0.295	0.768




```
##
## Call:
## lm(formula = price ~ mileage + IndWA + IndWA * mileage, data = true_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8893  -2.7472  -0.4353   3.6566  13.0413
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   36.059312   0.842970  42.777   <2e-16 ***
## mileage       -0.194625   0.010317 -18.864   <2e-16 ***
## IndWA          0.420312   1.538115   0.273    0.785
## mileage:IndWA -0.002542   0.018201  -0.140    0.889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.828 on 174 degrees of freedom
## Multiple R-squared:  0.7524, Adjusted R-squared:  0.7482
## F-statistic: 176.3 on 3 and 174 DF,  p-value: < 2.2e-16
```

Fitted model for (location 1):

$$\widehat{SeattlePrice} = -0.19(Mileage) + 36.35386$$

Fitted model for (location 2):

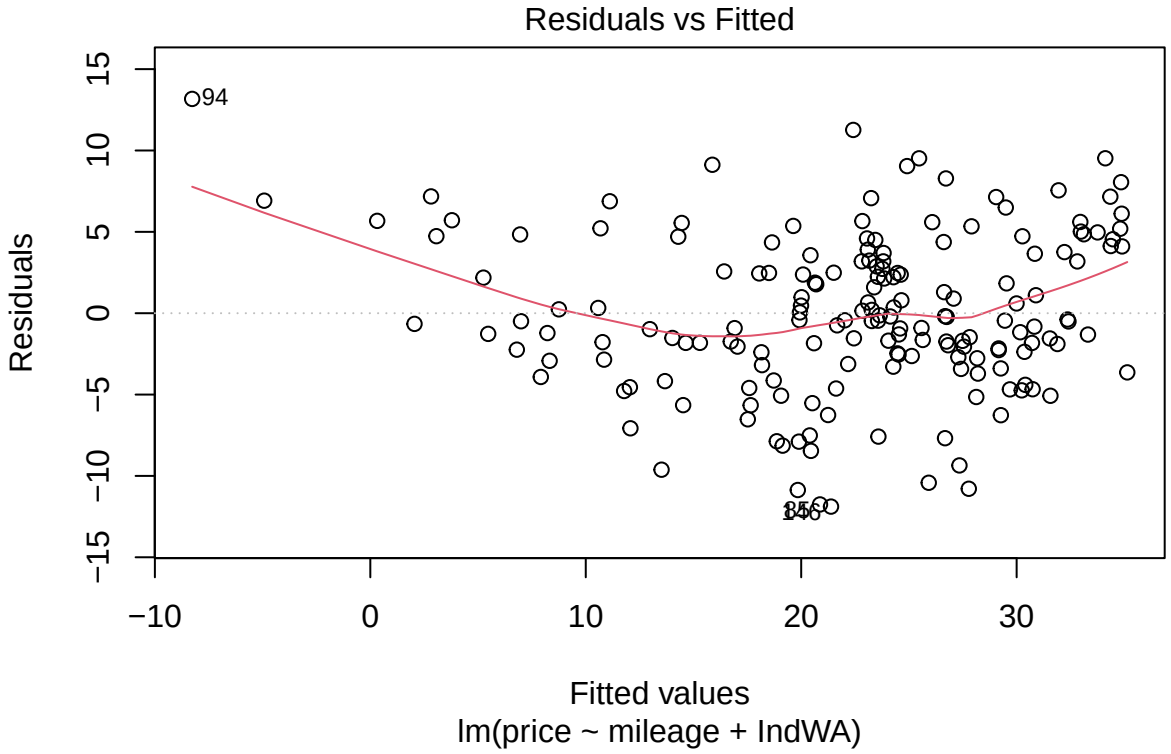
$$\widehat{ChicagoPrice} = -0.19(Mileage) + 36.116857$$

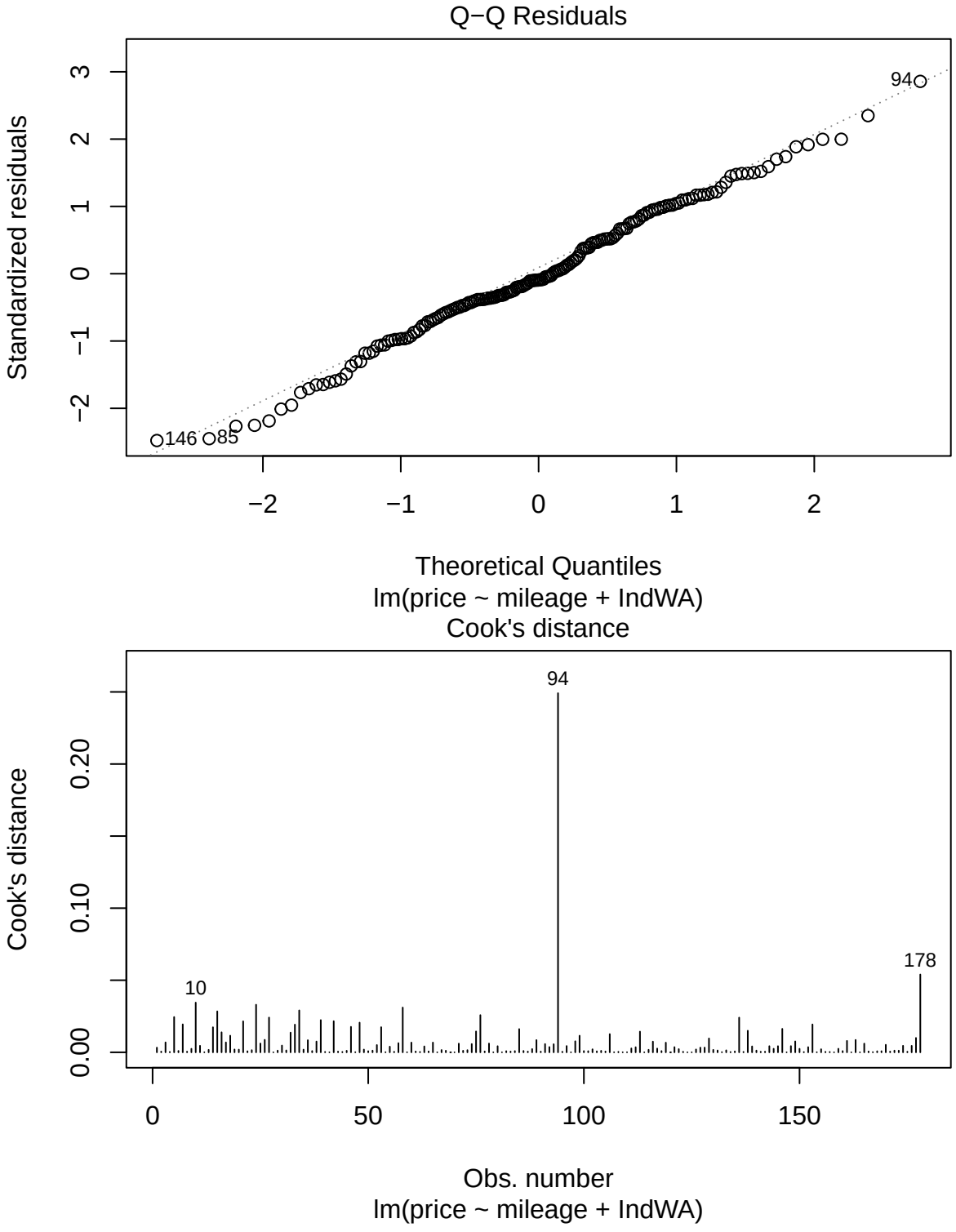
Comments

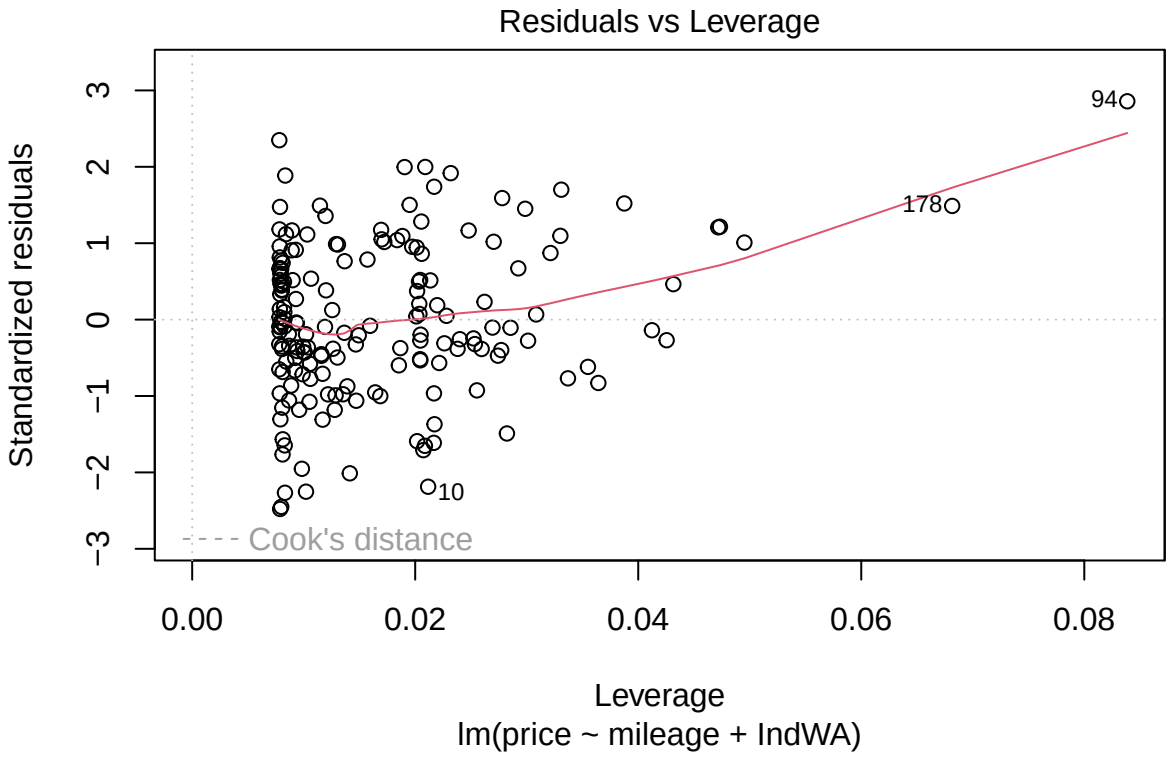
Between the two possible models, the lack of an interaction term resulted in a higher r^2 value. Having a higher r^2 with less regressors makes for a better model choice. Having a model that is dependent on less variables makes it more consistent and easier to manage. Normally our adjusted r^2 would go up whenever we add variables, but if it goes up when we have less variables we know that we have a much more efficient model. Even though the r^2 only increased a little bit, maybe even enough for it to be negligible, the reduction of an entire variable makes it a better choice.

Assess

Figures:




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ mileage + IndWA)
94
146
85
Cook's distance
Cook's distance
Obs. number
lm(price ~ mileage + IndWA)
94
178
10



Comments Surprisingly, it looks like all necessary conditions are met. The data is linear, the error looks quite normal, and the variance seems pretty equally distributed. My only real concern is the adjusted r^2 value for this model. With our residual and Q-Q plot, I would think that our adjusted r^2 value would be a bit higher. Hopefully in this next portion we can perform some adequate manipulation to improve the model.

Use

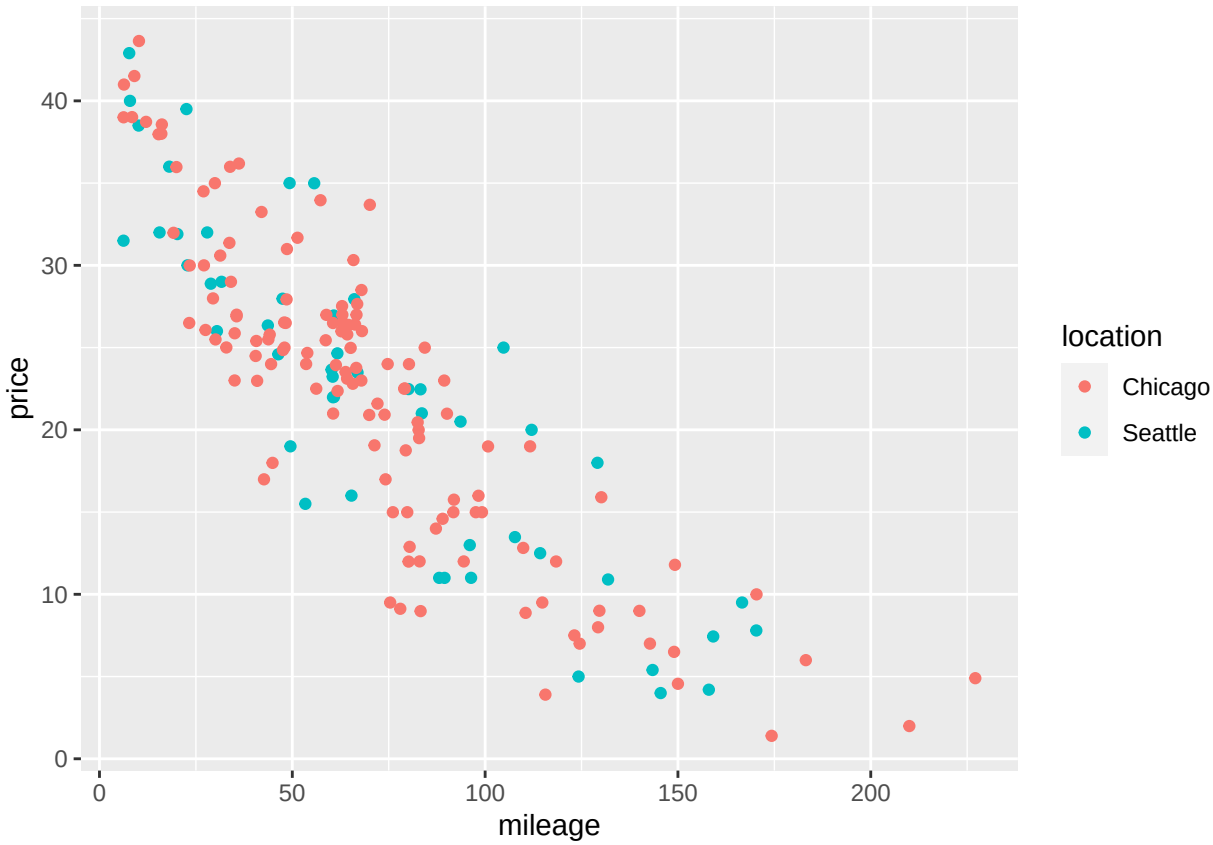
There is not enough evidence to suggest a statistically significant difference in price of a Nissan Maxima between Seattle and Chicago. Even though our data shows a difference in price of about $237, our p-value is not nearly low enough to be able to claim that this difference is statistically significant.

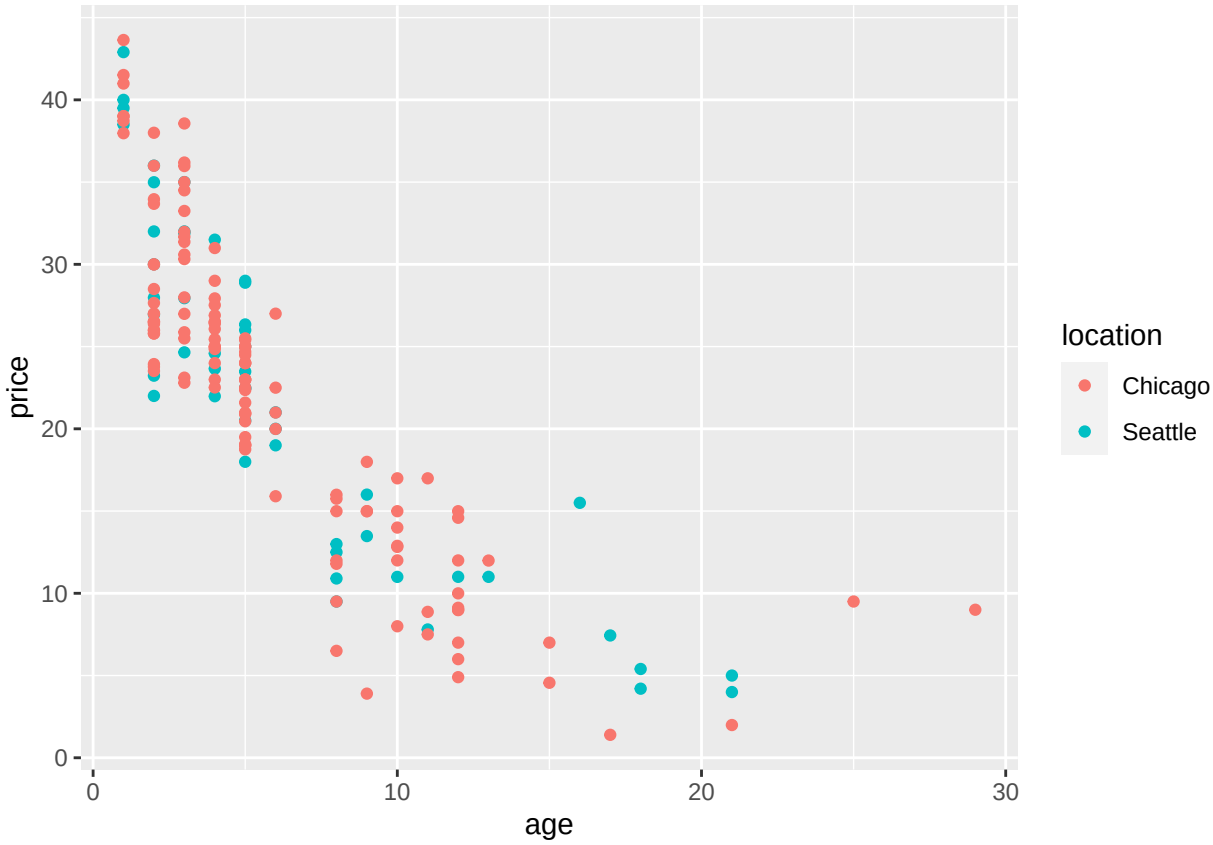
Research Question 2

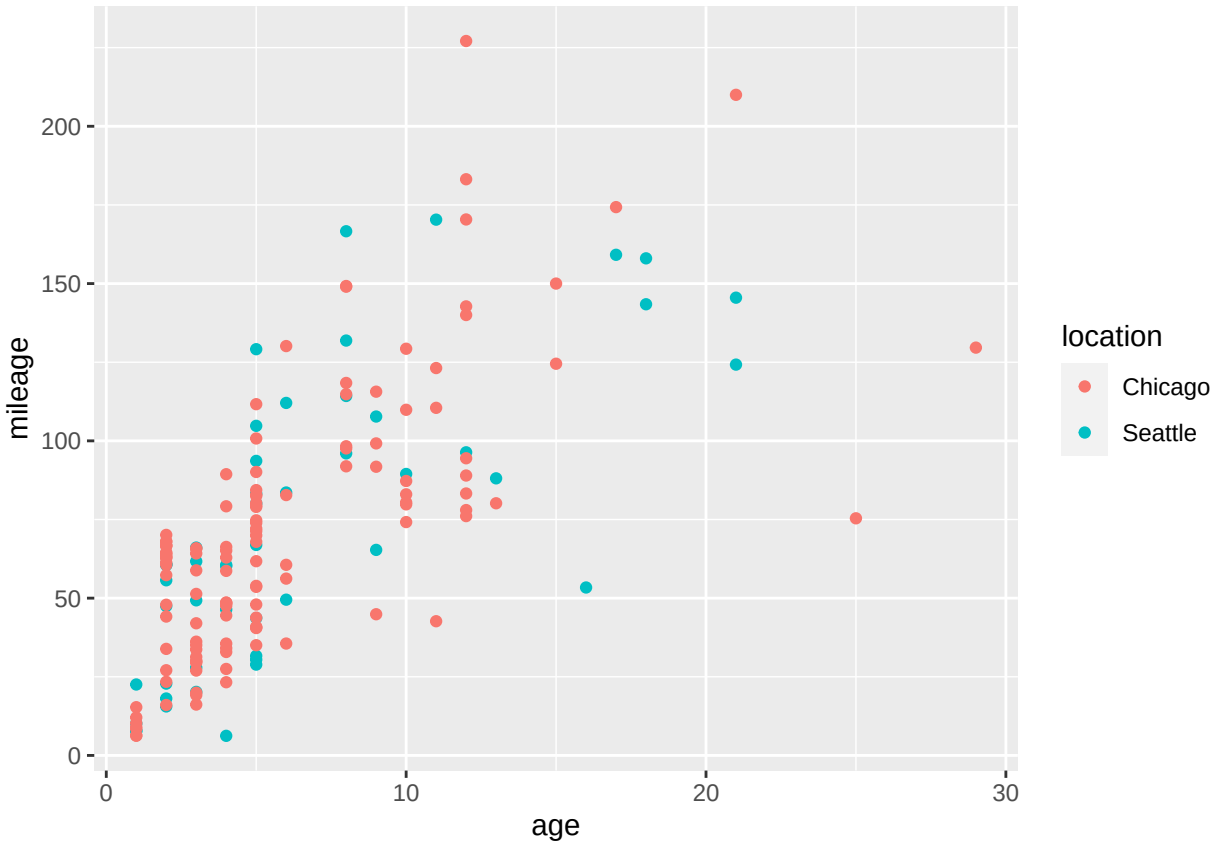
After accounting for a linear relationship between age and price and between mileage and price, is there a difference in price between the locations?

Exploratory data analysis

Figures:


price
mileage
0
10
20
30
40
0
50
100
150
200
location
Chicago
Seattle


price
age
0
10
20
30
0
10
20
30
location
Chicago
Seattle



Comments

It still looks like the relationship between price and mileage is quite linear. When comparing price and age as well as age and mileage, the relationship looks like it is no longer linear. Price and age looks like it has a negative exponential decay, while age and mileage looks like it has a bit of a "cornucopia" effect. As the age increases the variance in mileage increases dramatically. Both of these issues can potentially be solved with a logistic or exponential transformation.

Model fitting

```
##
## Call:
## lm(formula = price ~ mileage + IndWA + log(age), data = true_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9852 -2.0157 -0.0134  1.6290  7.6607
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.541819   0.525016  77.220   <2e-16 ***
## mileage     -0.100594   0.007709 -13.049   <2e-16 ***
## IndWA        0.291007   0.497450   0.585    0.559
## log(age)    -7.293184   0.434120 -16.800   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.982 on 174 degrees of freedom
```

```
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.9039
## F-statistic: 556.2 on 3 and 174 DF,  p-value: < 2.2e-16
```

SUMMARY TABLE 3:

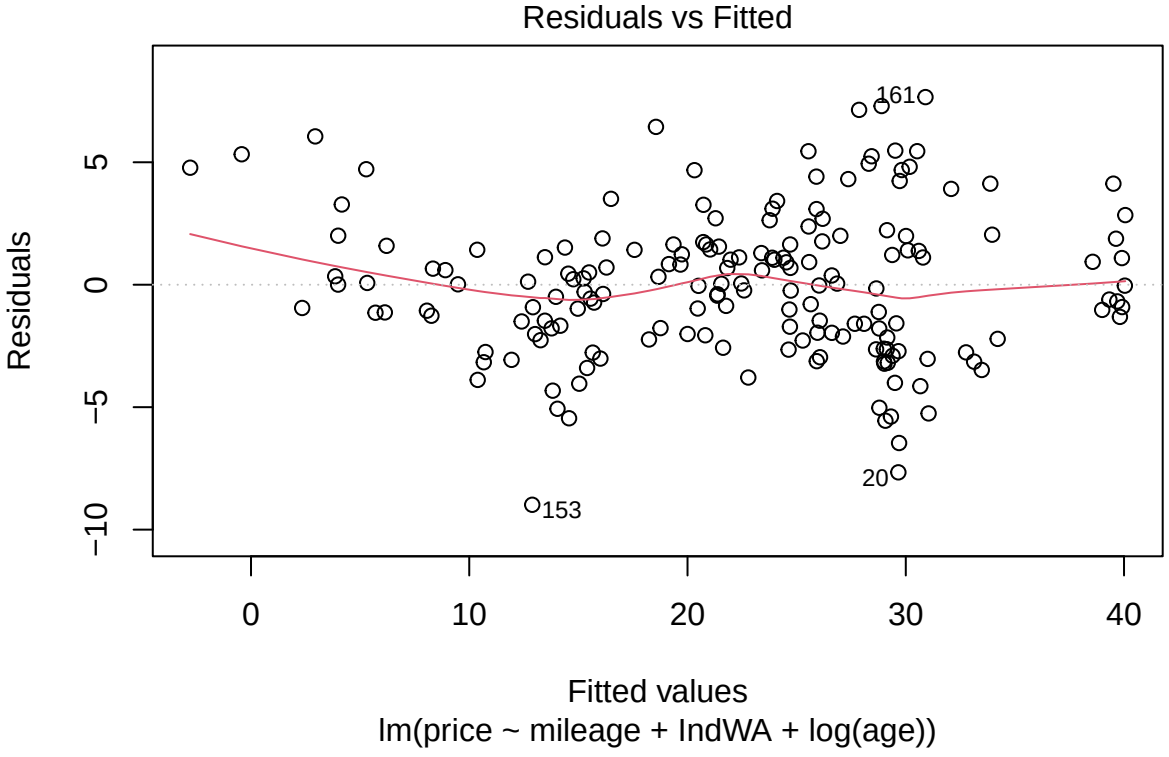
	estimate	test-statistic	p-value
intercept	40.54	77.22	<2e-16
mileage	-0.10	-13.049	<2e-16
log(age)	-7.29	-16.80	<2e-16
location	0.29	0.58	0.559

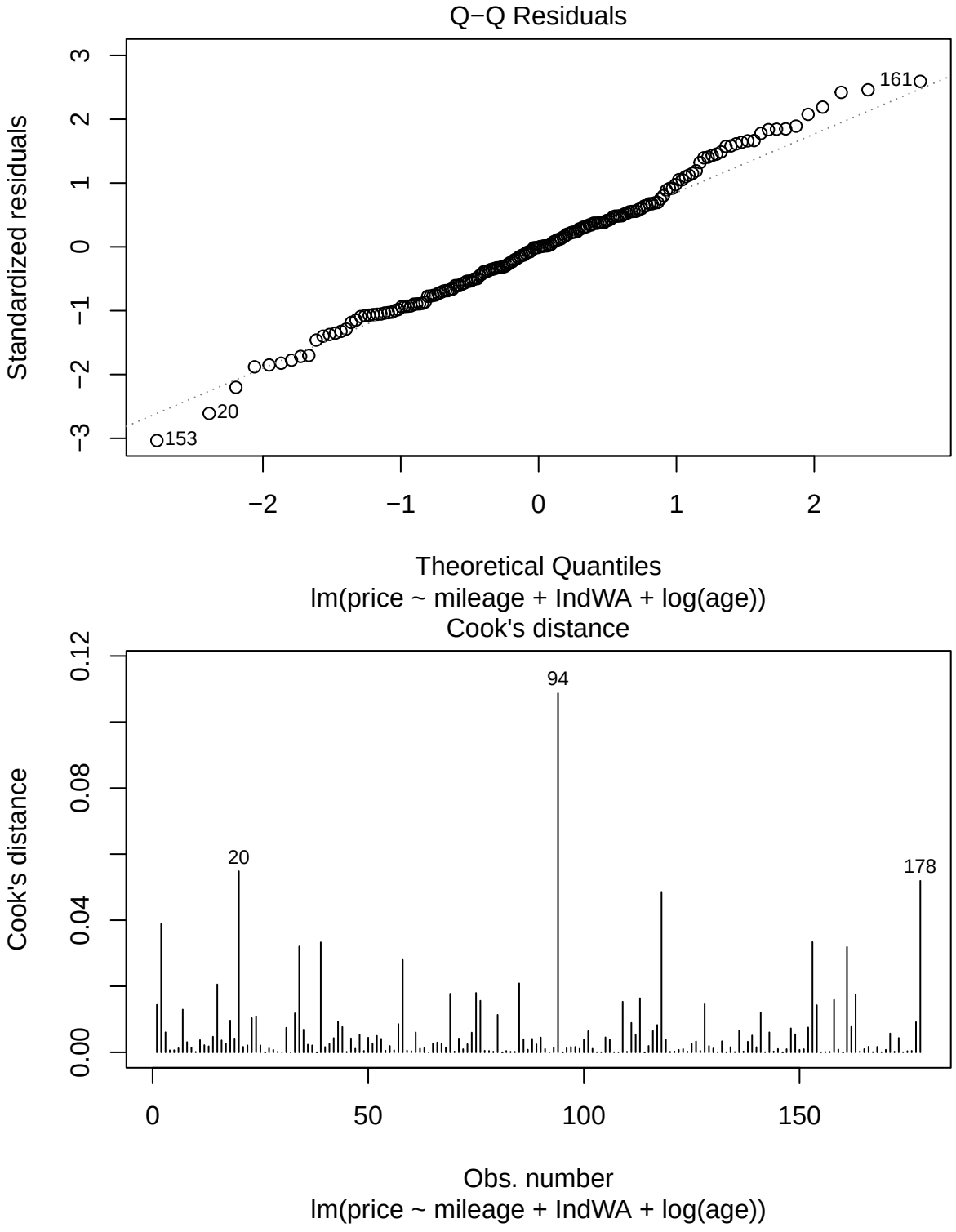
Interpretations in context

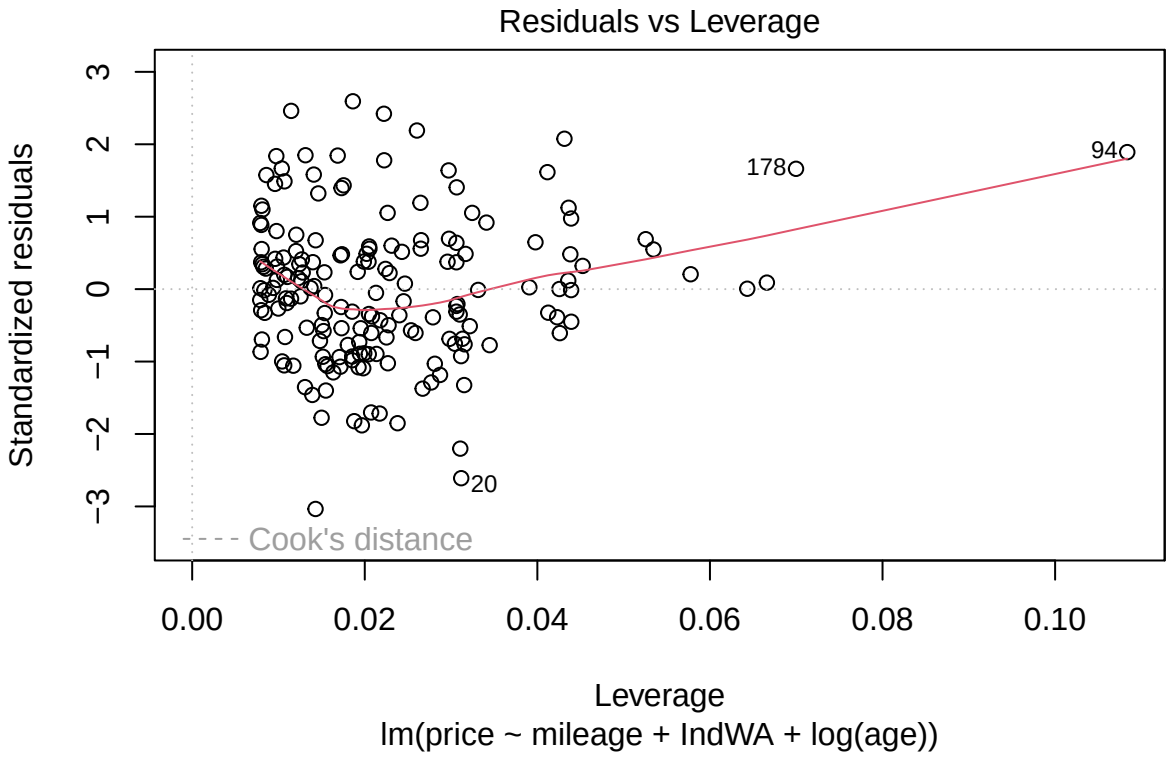
- intercept: We estimate the price of a brand new Nissan Maxima to be 40,540 dollars.
- mileage: For every 1,000 miles, holding other variables constant, the price of a Maxima decreases by an average of 100 dollars.
- age: For every log(year), while all other variables are constant, the price of a Maxima decreases by an average of 7,290 dollars.
- location: With not enough statistical significance to make a confident claim, there is a difference of 290 dollars when all other variables are constant.

Assess

Figures:




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ mileage + IndWA + log(age))
161
20
153
Cook's distance
Cook's distance
Obs. number
lm(price ~ mileage + IndWA + log(age))
94
20
178



Comments

The conditions seem to be very well met with this model. The data seems to be quite linear, the error appears to be much more normal than our previous model, and the variance appears to be quite equal. Although some points are marked as outliers, the highest Cook's Distance value of any point is 0.11, which is very encouraging. I have no concerns with this model whatsoever, but I am curious if it can still be improved.

Use

There still doesn't not appear to be a significant difference in price of Nissan Maximas between Seattle and Chicago. Although the estimated difference has increased with this model by 60 dollars, the p-value has decreased. That being said, the p-value is still above 0.5, so we do not have nearly enough evidence to suggest or claim that there is a statistically significant difference in price between Seattle and Chicago.

Research question 3

What is the best model for predicting price using the variables available?

Choose

Final fitted model:

$$\widehat{Price} = -3.44log(Mileage) + 1.37log(Age) - 2.075[log(Mileage) * log(Age)] + 47.8734$$

Summary of final model:

```
##
## Call:
## lm(formula = price ~ log(mileage) + IndWA + log(age) + log(mileage) *
##     log(age), data = true_data)
```

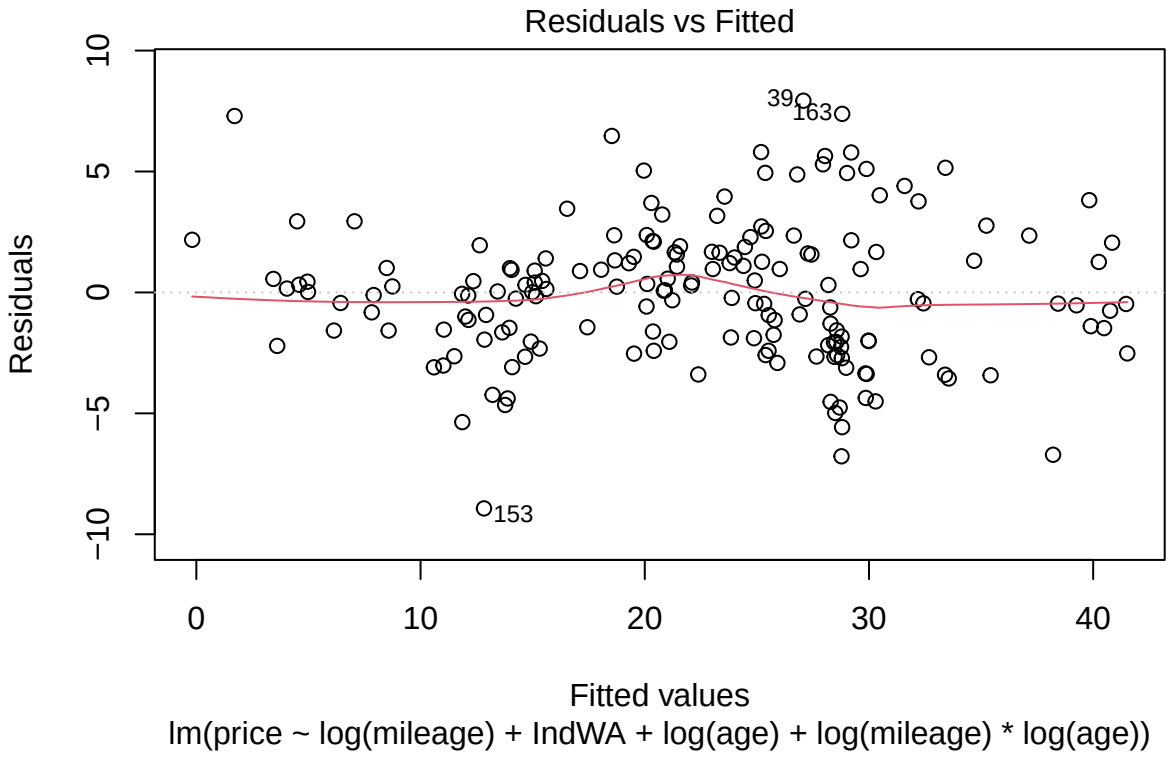
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9296 -2.0064  0.0133  1.6328  7.9256
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           47.83127    1.91454  24.983  < 2e-16 ***
## log(mileage)          -3.44211    0.54614  -6.303 2.35e-09 ***
## IndWA                  0.04228    0.48577   0.087    0.931
## log(age)               1.37082    1.49676   0.916    0.361
## log(mileage):log(age) -2.07538    0.33422  -6.210 3.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.899 on 173 degrees of freedom
## Multiple R-squared:  0.9112, Adjusted R-squared:  0.9092
## F-statistic:   444 on 4 and 173 DF,  p-value: < 2.2e-16
```

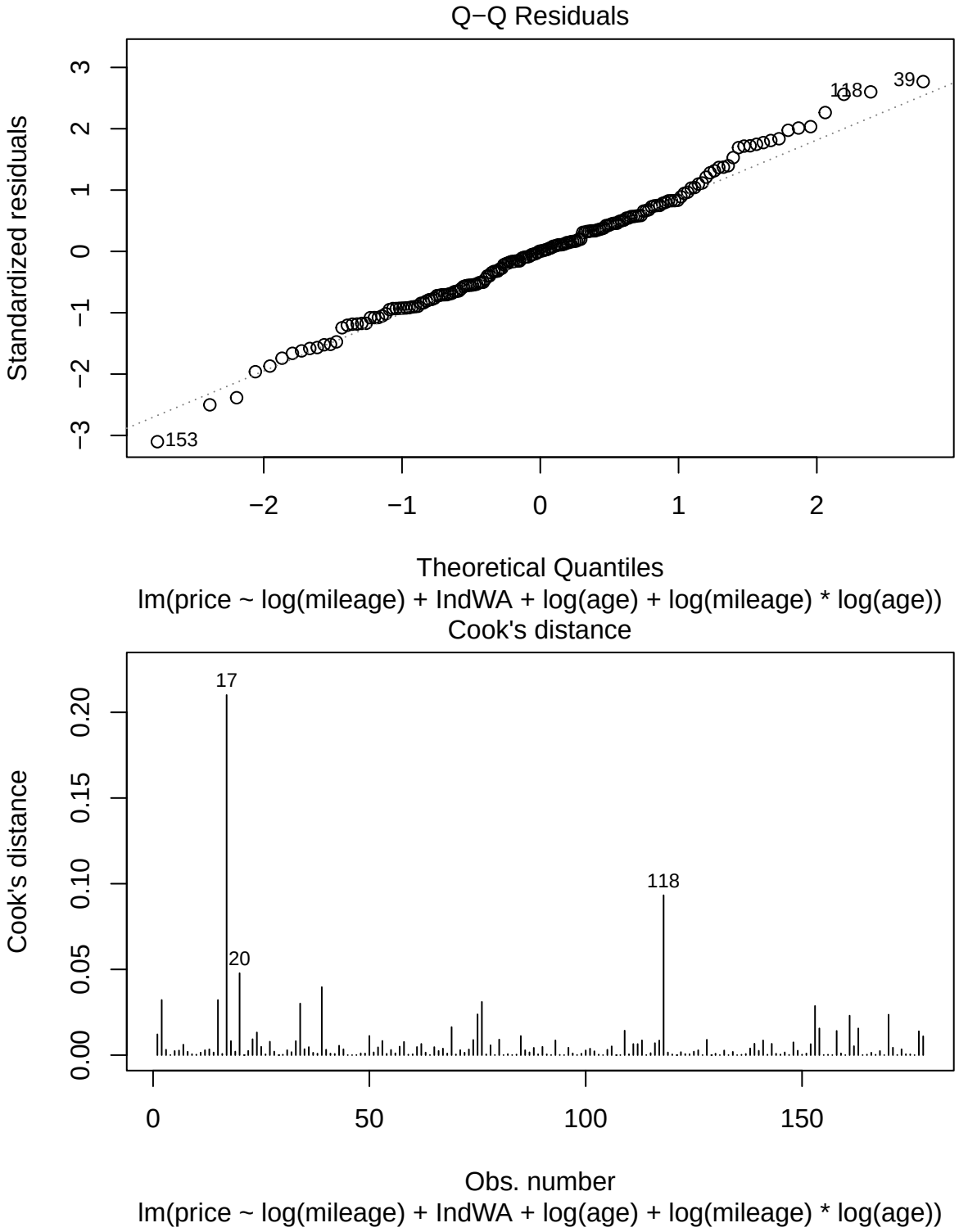
Comments

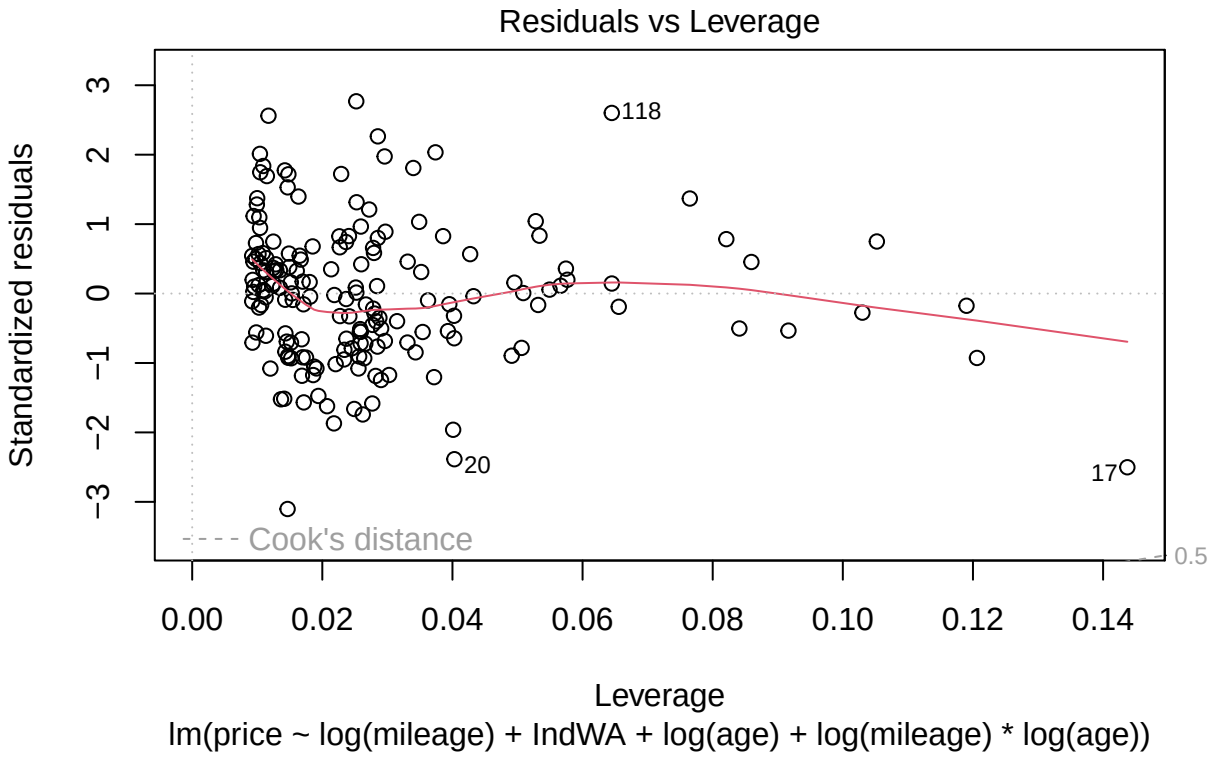
I decided on this final model after a lot of trial and error along with some manipulation. Some of the variables positively impacted my adjusted r^2 value just by applying a logarithm, so I tried this with all of my variables, and even creating an interaction term out of them. This worked out quite well, with a final r^2 value of 0.9092. Although two of the variables have pretty high p-values, I'm much more satisfied with having a r^2 that is this high. Removing the variables with high p-values seems to negatively impact this model, so I believe their presence to be beneficial here.

Assess

Figures:




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ log(mileage) + IndWA + log(age) + log(mileage) * log(age))
Cook's distance
Cook's distance
Obs. number
lm(price ~ log(mileage) + IndWA + log(age) + log(mileage) * log(age))



Comments

All necessary conditions are met with this model. The data is linear, the error is normally distributed, and variance is equal throughout the model. I believe many of the variables in this model to be "almost" linear but not quite, which is why applying logarithms seems to be so effective. For these reasons, as well as I very high r^2 value, this model seems to capture the data quite accurately.

Use

Location: Seattle

Mileage = 40,000 Age = 3 years Price = -3.44log(mileage) + 1.37log(age) - 2.075[log(mileage)*log(age)] + 47.8734

```
##        fit      lwr      upr
## 1 28.27125 22.47913 34.06337
```

We predict a Nissan Maxima in Seattle that is 3 years old with 40,000 miles to be between 22,479 dollars and 34,063 dollars with 95% confidence. The predicted price is 28,271 dollars.

Conclusion

In this project I learned that the difference in prices of Nissan Maximas between Chicago and Seattle is negligible. That being said, the price in Seattle is projected to be slightly higher. From the summary statistics in my final model, it seems that mileage was the biggest factor in price. We know this because it has the largest negative coefficient within our model summary. This is also in line with the previous models created in this project. Given the adjusted r^2 value of this model, I think it is a good one, but I think it can also be better. If we were able include any information about vehicle history, like the amount of money spent on repairs, or whether or not it was owned by a smoker, I believe the model could be improved quite a bit.