MATH 324: Multiple Linear Regression project

Ellise Putnam

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.

Introduction

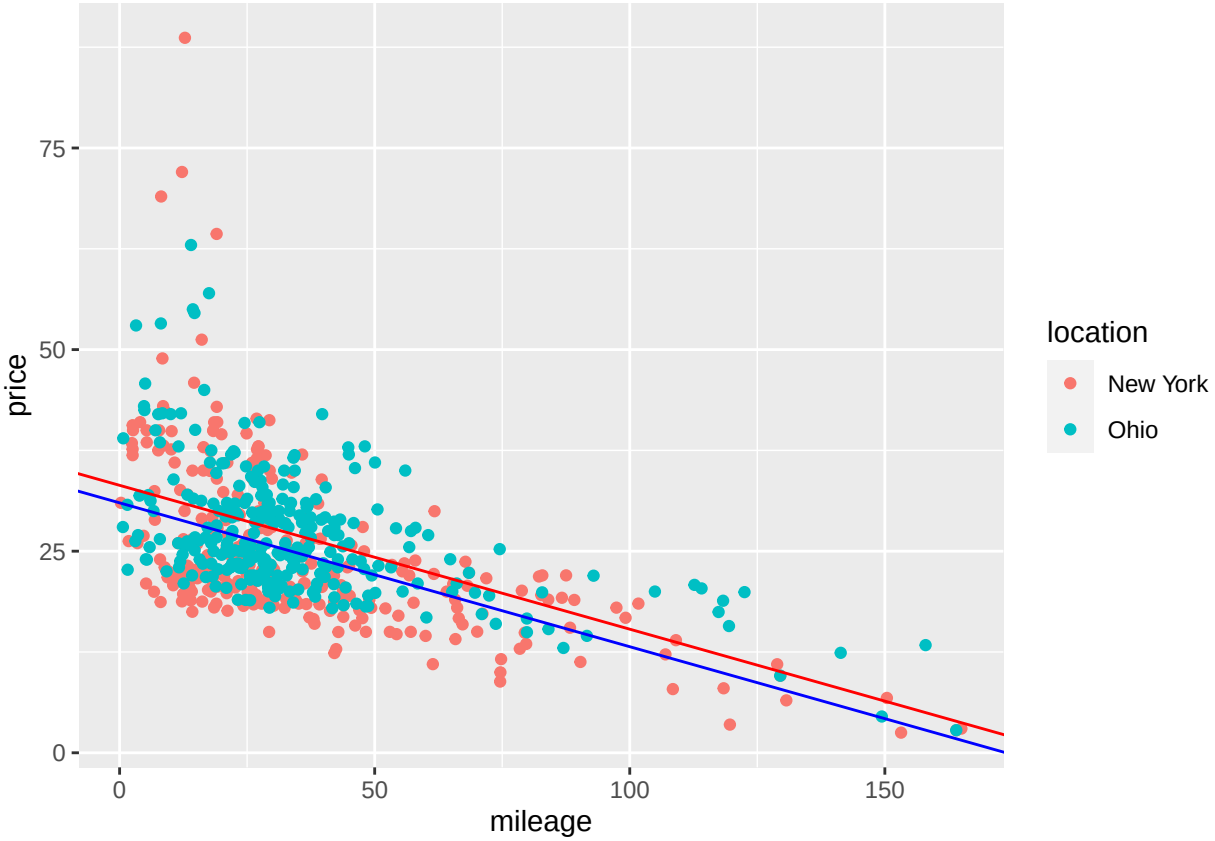
For this multiple linear regression project, I decided to look into the prices and mileage of used **Jeep Wranglers**. I did a little research and found that the Jeep headquarters are in **Toledo, Ohio**. Additionally, Google says that the state with the most jeep owners is **New York**. I thought it would be interesting to compare the two locations because one state holds the headquarters of Jeep while the other holds the most drivers of jeeps. In theory, New York would have cars with less mileage because there is more public transportation, and I would think that Ohio would have older cars because the headquarters are there.

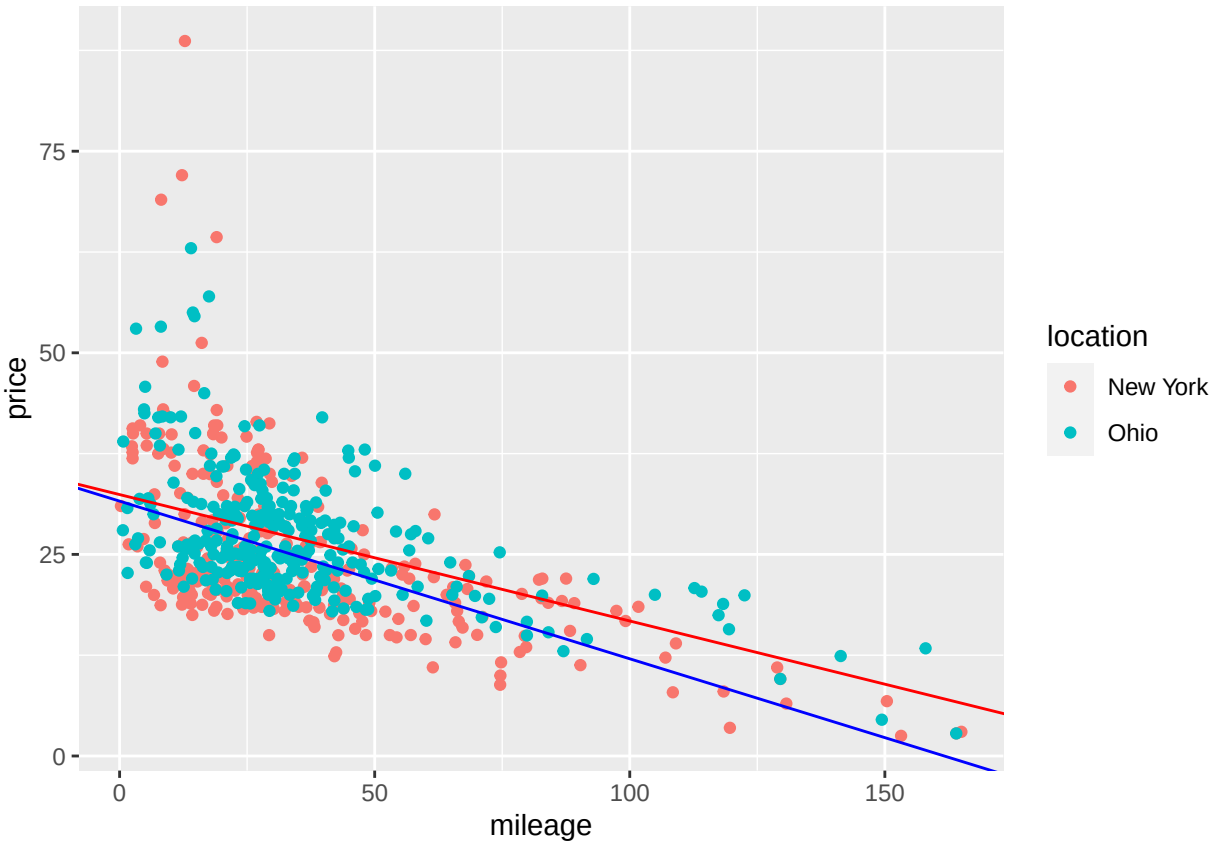
Research question 1

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
New York	300	24.63	10.10444	35.687	28.86603
Ohio	300	26.99	7.758758	34.63	25.66931

Comments:

After looking at these statistics, the average price for a used car is greater in Ohio than in New York where Ohio is $26,990 and New York is $24,630. At first this was surprising to me because I would think New York would have more expensive prices, however, the more expensive Wranglers were in Ohio though only by $2,000. It is important to note that New York does have the most expensive (max) car but on average, has less expensive used cars. Additionally, I found it interesting that New York also, on average, had a greater mileage per car, which was also contrary to what I would have thought considering the plethora of public transportation in New York. Additionally, looking at the scatterplot, we can see that the price intercept is roughly the same for both locations, however, New York has a steeper slope, meaning that as the price decreases the mileage increases at a greater rate in New York than in Ohio, or that New York cars lose value in price faster based on the increase of mileage as compared to Ohio cars. If I were to answer the research question now without diving any deeper into the data, I would say that there is a very minimal difference, if not any, between prices of Jeep Wranglers based on the location.

Model fitting

MODEL SUMMARY TABLE 1: (same slope different intercepts)

	estimate	test-statistic	p-value
intercept	30.99412	51.96	2e-16
mileage	-0.17840	-15.71	2e-16
locationOhio	2.17357	3.51	0.000482

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

	estimate	test-statistic	p-value
intercept	31.60503	45.427	<2e-16
mileage	-0.19552	-12.891	<2e-16
locationOhio	0.81301	0.803	0.4220
mileage:locationOhio	0.03877	1.698	0.0899

Fitted model for (location 1, New York):

$\hat{price}$ = 31.60503 - 0.19552 * mileage

Fitted model for (location 2, Ohio):

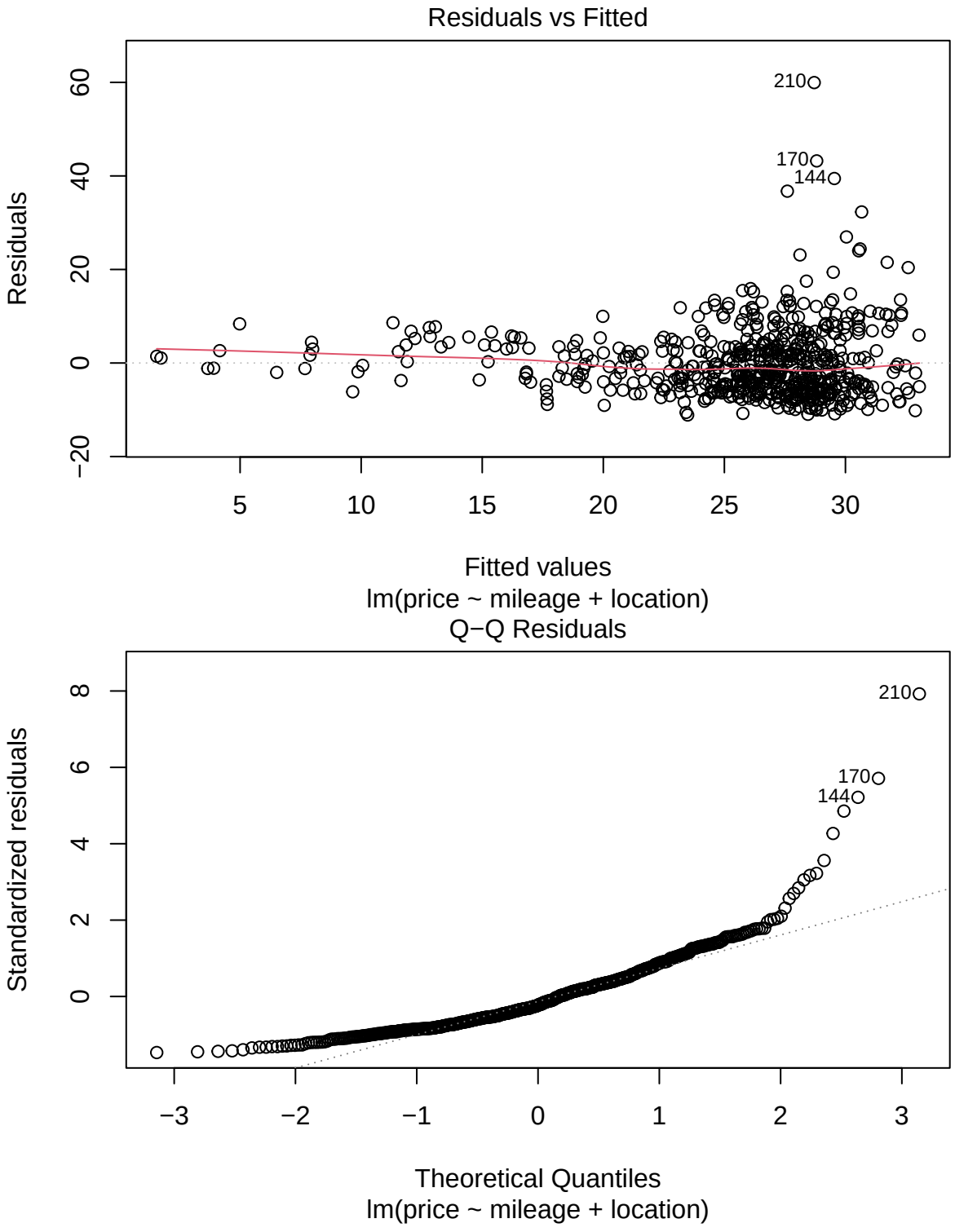
$\hat{price}$ = 31.6438 - 0.23429 * mileage

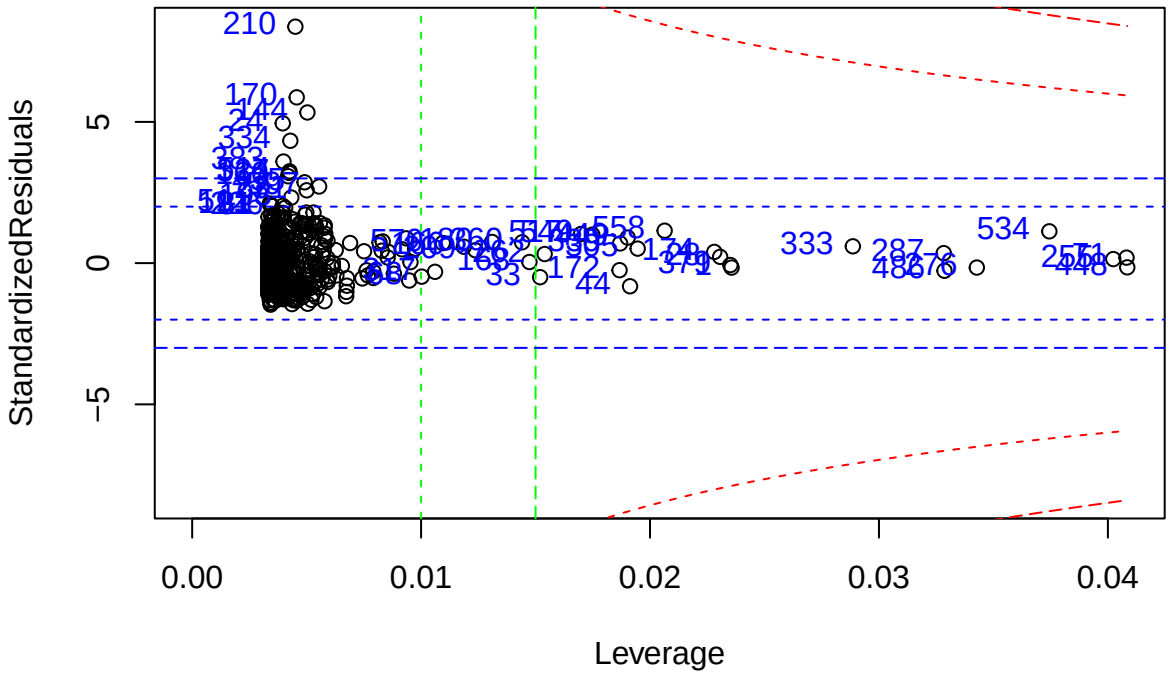
Comments

After looking at the final equations for New York and Ohio, the intercepts and slope are not significantly different. The difference in intercepts is 0.03877 and the difference in slopes is 0.03877. Additionally, looking at the r-squared values for both models we see $r^2 = 0.3046$ for the first model and $r^2 = 0.308$ for the second model. These numbers tell us that for both models, 30% of the variability in price can be explained by the models, and therefor there is no significant difference in the models because their r-squared values are so close. As a result of these models being so similar, I decided to choose the first one, same slope and different intercepts, because that model is simpler and therefor will be better for interpreting and understanding.

Assess

Figures:


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ mileage + location)
210
170
144
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ mileage + location)



Comments

By looking at our two plots, residuals vs. fitted and our Q-Q plot, we can now assess our LNE conditions. First off, in our residuals vs. fitted plot we have a classic cornucopia flaring plot which tells us that our data is not linearity or equal in variance. Additionally, while our Q-Q plot looks fairly normal, we do see skewing towards the beginning and end of the residual line and therefor this data is not normal either. In order to solve this we can use a transformation in hope of achieving our LNE conditions. Additionally, our cooks plot, while having a few outliers as mentioned, looks good in terms of no points falling outside of the red dotted lines, or having high values of residuals.

Use

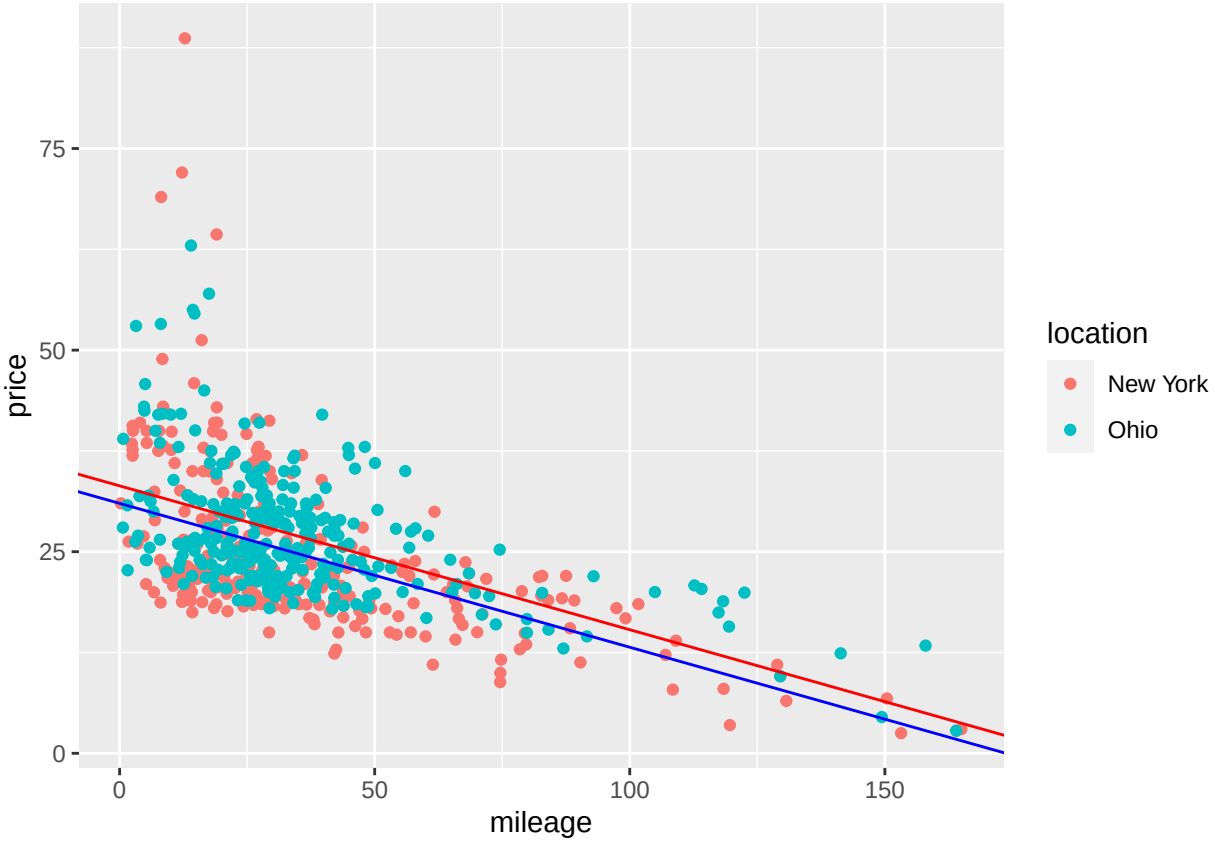
In conclusion, and after deciding to use our first model where we have the same slopes and different intercepts, and after accounting for mileage, there is a difference in price based on location because we have a small enough p-value to reject the null in favor of the alternative. Additionally, this tells us that, on average, Jeep Wranglers in Ohio are predicted to have a higher prices compared to those in New York when assuming a linear relationship with mileage.

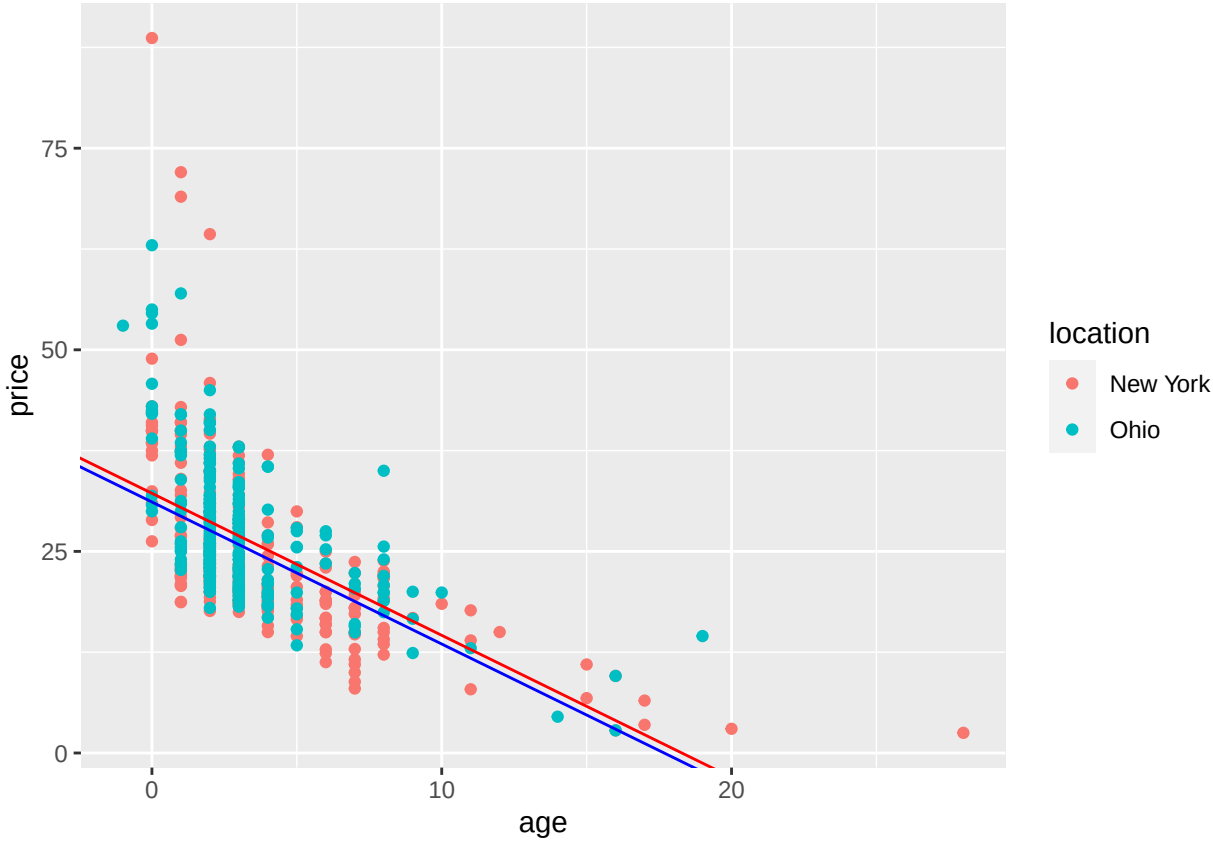
Research Question 2

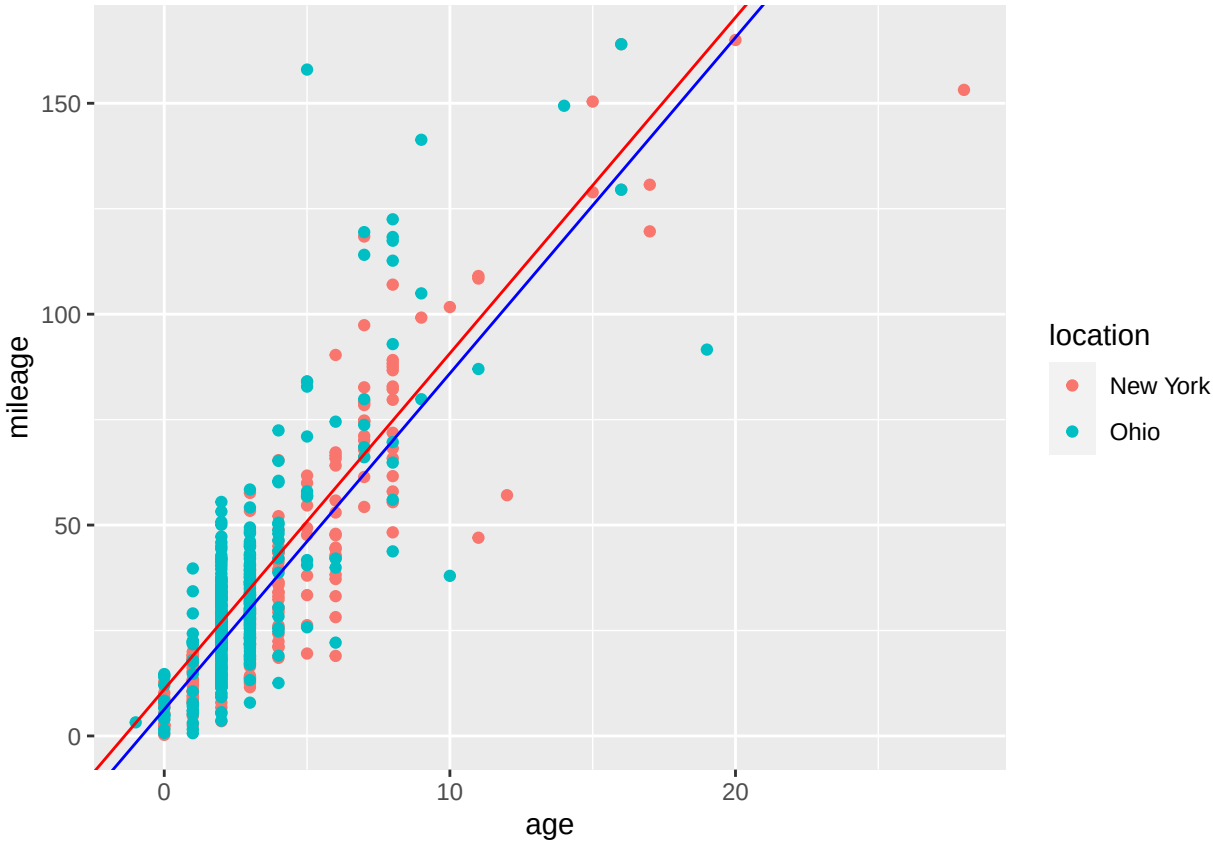
After accounting for a linear relationship between age and price and between mileage and price, is there a difference in price between the locations?

Exploratory data analysis

Figures:




price
75
50
25
0
0
10
20
age
location
New York
Ohio


mileage
150
100
50
0
0
10
20
age
location
New York
Ohio

Comments

After looking at the plots we can see that all three have a linear relationship, while some stronger than others. First, our scatterplot with price and mileage demonstrates a strong negative linear relationship, that is, as mileage increases the price of the car decreases. The next scatterplot demonstrates the negative linear relationship between price and age, showing that as the older a car gets in age, the price decreases, though not as strong of a negative relationship as mileage and price. Lastly, we have a positive linear relationship between mileage and age, that is, as the age of a car increases the mileage on a car tends to increase as well. Though there are some outliers presented in all 3 plots, the trends demonstrate linear relationships.

Model fitting

SUMMARY TABLE 3:

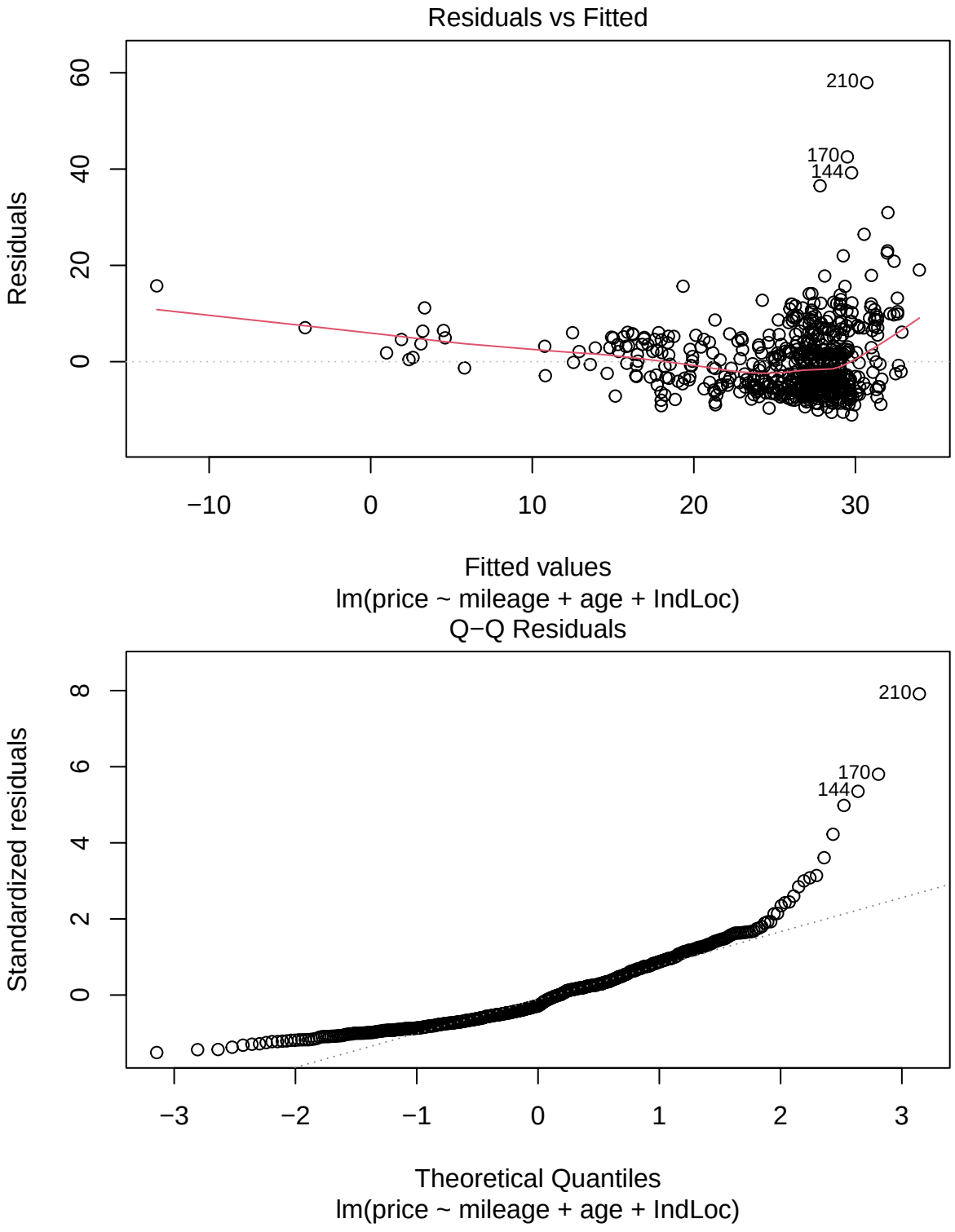
	estimate	test-statistic	p-value
intercept	31.53469	53.984	$<$ 2e-16
mileage	-0.06491	-3.086	0.00213
age	-1.24415	-6.330	4.83e-10
locationOhio	1.38544	2.261	0.02411

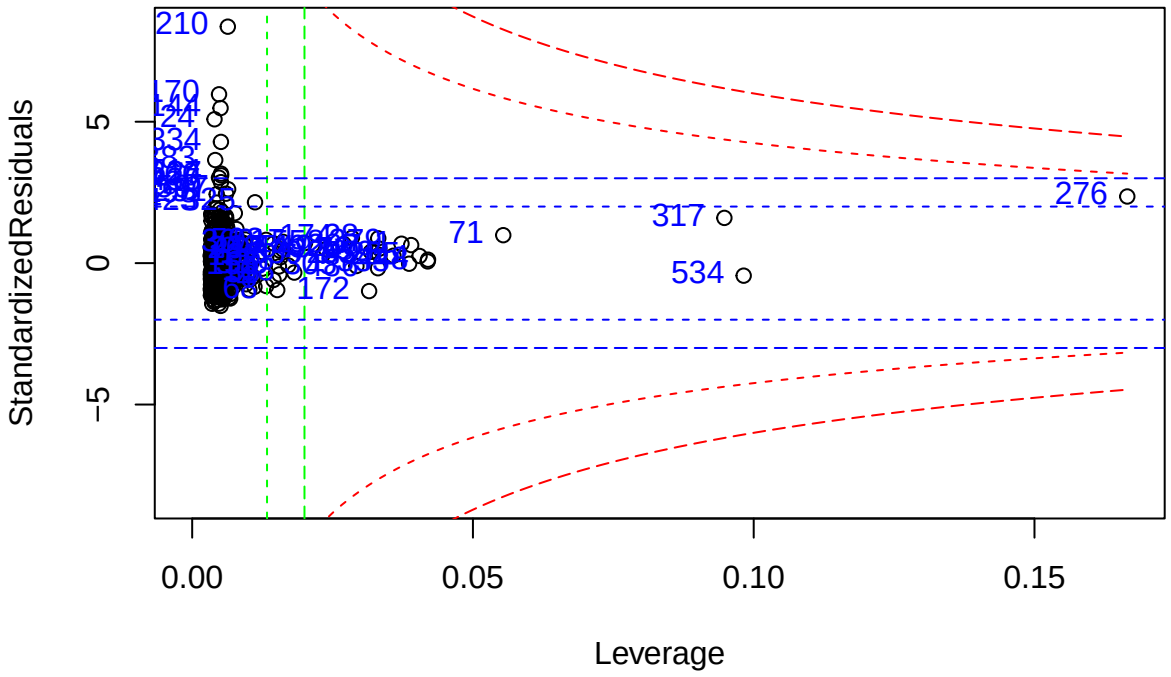
Interpretations in context

- intercept: After accounting for mileage, age, and location, the predicted price of a Jeep Wrangler is $31,534.69.
- mileage: Keeping location and age constant, the predicted price of a car decreases by $64.91 for each 1000 miles.
- age: Keeping location and mileage constant, the predicted price of a car decreases by $1244.15 for every year older it is.
- location: After accounting for mileage and age, a Jeep Wrangler in Ohio is predicted to cost $1385.44 more than in New York.

Assess

Figures:


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ mileage + age + IndLoc)
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ mileage + age + IndLoc)
210
170
144

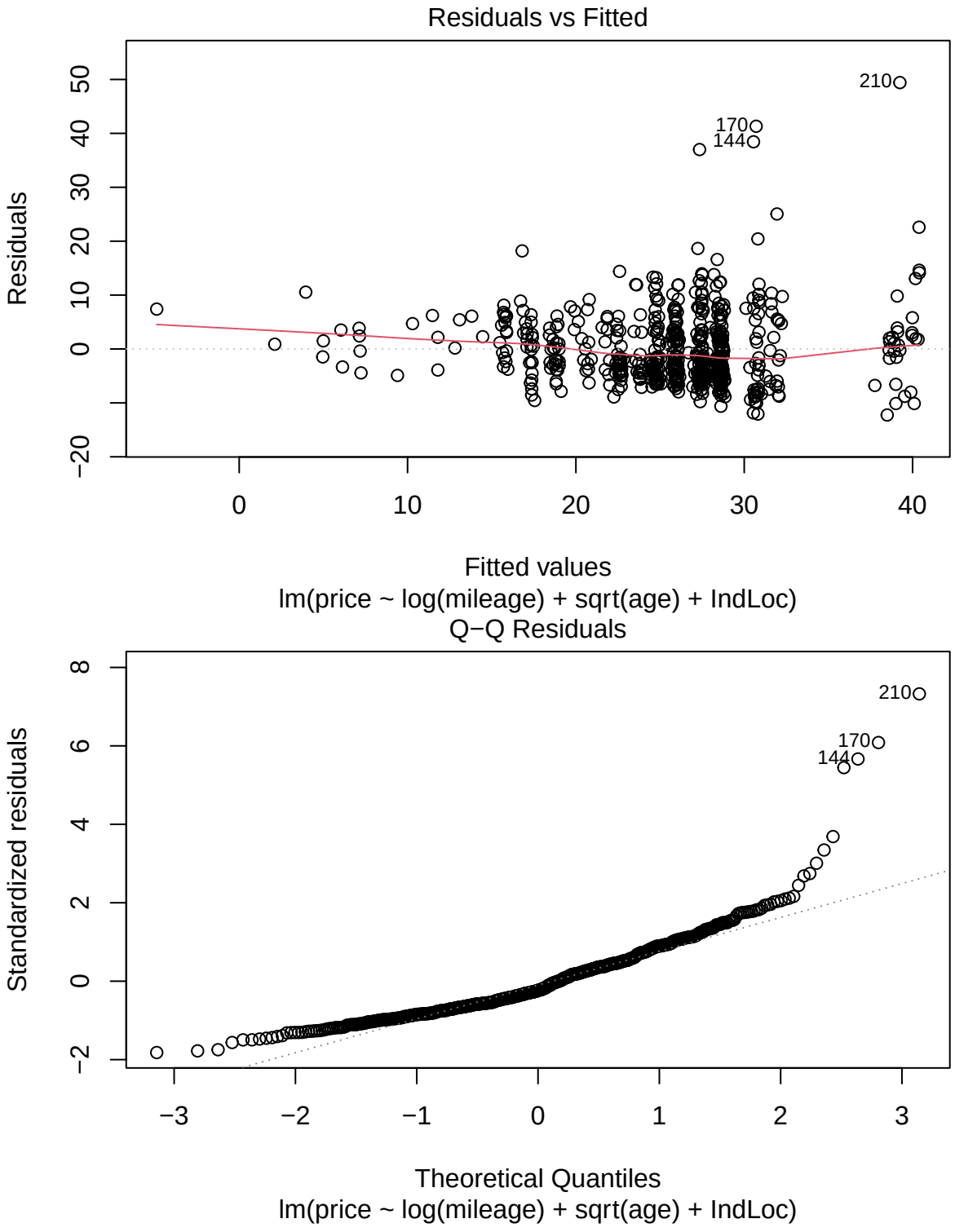


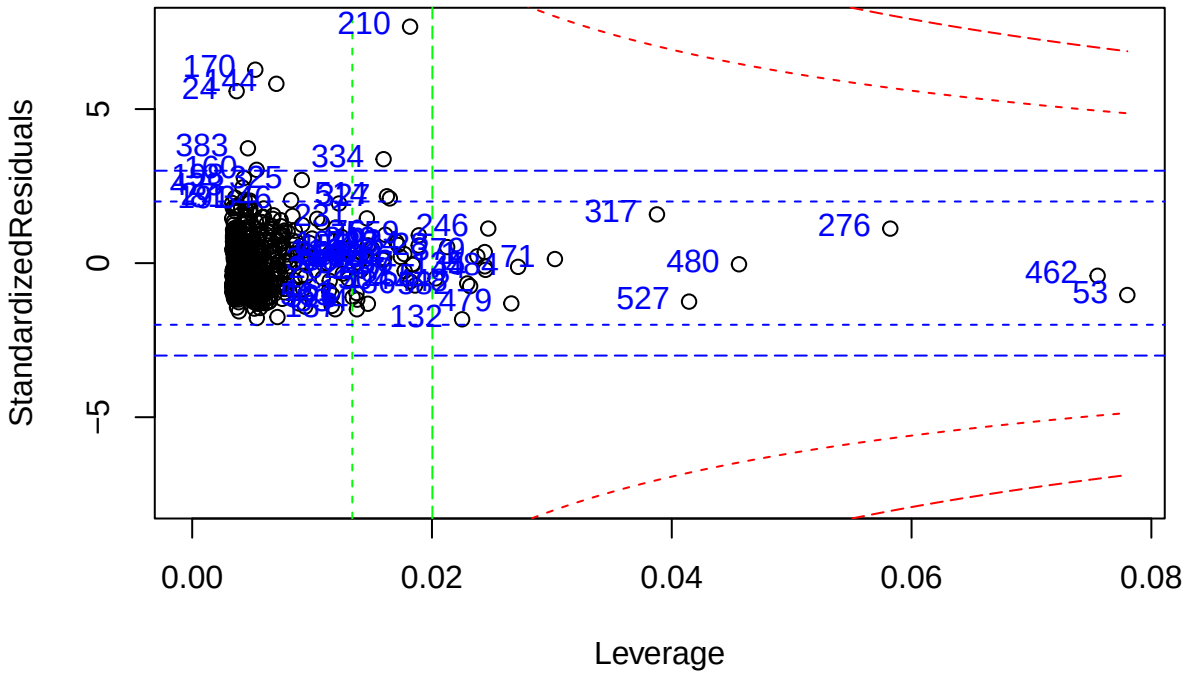
Comments

Similar to the plots in Research Question 1, these graphs are not good in terms of meeting LNE conditions. In terms of equal variance, we have an upwards flaring trend that suggests unequal variance. Additionally, we have some skewing from our residual line in our Q-Q plot which suggests a lack of normality, and lastly in our residuals vs. fitted plot we can see that we could argue linearity if we removed some outliers. There appears to be an equal spread of points above and below our horizontal line when not considering the outliers. In terms of our cooks plot, I notice a few things that might be helpful to note. We have a few points that are numbered that demonstrate a high residual value which suggests a high influence on our multiple regression model. Additionally, we have a point or two with high leverage to the right side which also suggests points with high influence on our model. Overall, we have some points with high influence but they are within our red lines.

Transformation

```
## Warning in sqrt(age): NaNs produced
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ log(mileage) + sqrt(age) + IndLoc)
210
170
144
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ log(mileage) + sqrt(age) + IndLoc)




```
##
## Call:
## lm(formula = price ~ log(mileage) + sqrt(age) + IndLoc, data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.243  -4.621  -1.628   3.280  49.418
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.2689     1.2569  30.447   <2e-16 ***
## log(mileage)   0.3840     0.5674   0.677    0.499
## sqrt(age)     -8.5231     0.6378 -13.364   <2e-16 ***
## IndLoc         1.1011     0.5656   1.947    0.052 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.808 on 595 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4329, Adjusted R-squared:   0.43
## F-statistic: 151.4 on 3 and 595 DF,  p-value: < 2.2e-16
```

Comments After deciding to perform a few transformations on my model (price ~ log(mileage) + sqrt(age) + IndLoc), I found a better model to predict price that met the LNE conditions and provided a better Cooks plot.

Use

In conclusion, after using our model provided the transformations, after adjusting for mileage and age, and creating an indicator variable for location, there is not enough evidence to reject the null that there is no difference in price based on location, that is, we fail to reject the null. This is due to the insignificant p-values

for log(mileage) and IndLoc, while there could be a difference in price based on location, at this time we do not have enough evidence to say otherwise.

Research question 3

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

Choose

Final fitted model:

$\hat{\text{price}} = 30.99412 - 0.17840 * \text{mileage}^2 + 2.17357 * \text{IndLoc}$

summary of final model:

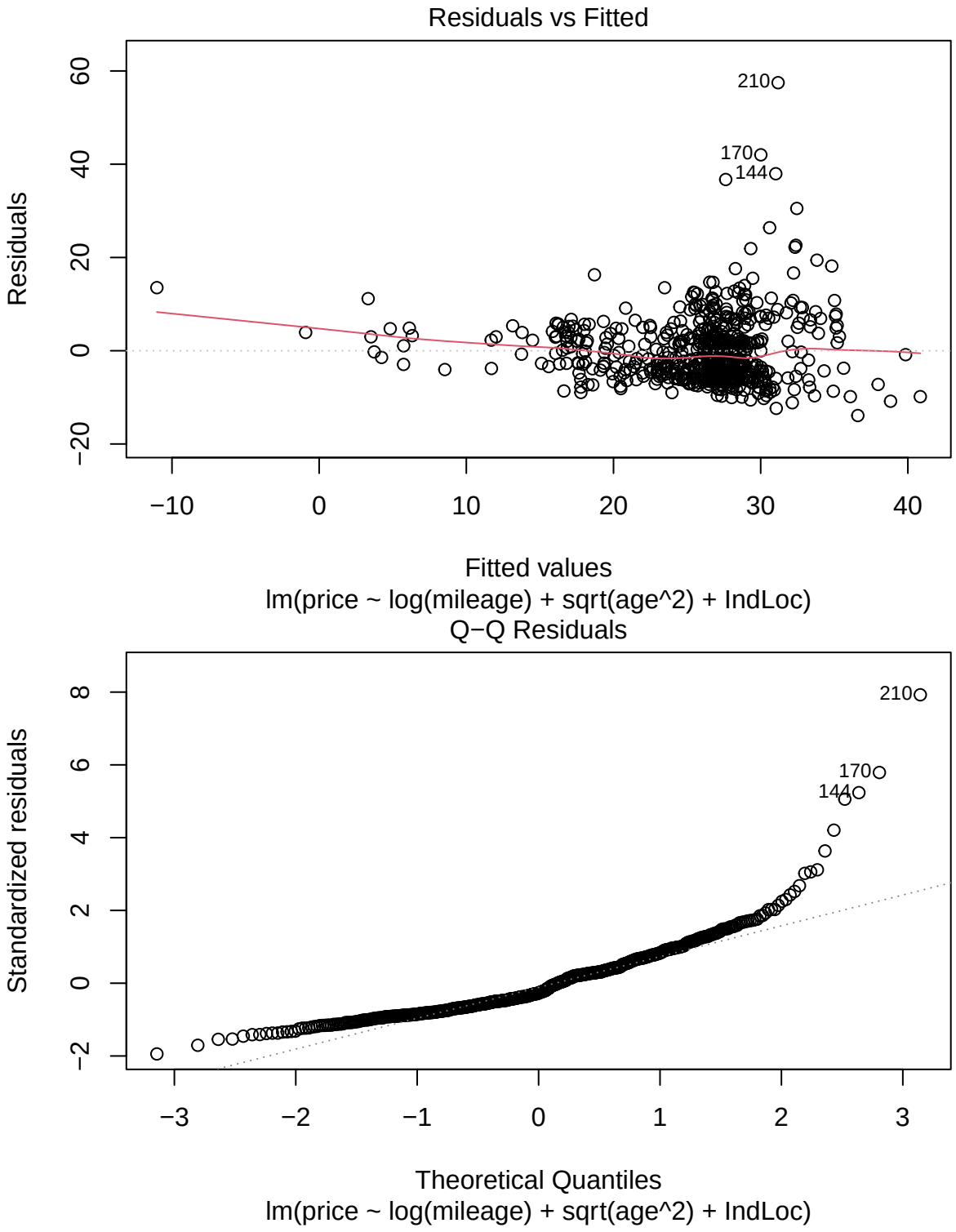
```
##
## Call:
## lm(formula = price ~ mileage + IndLoc + age^2, data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.073  -5.269  -2.055   3.548  57.963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.53469    0.58415  53.984  < 2e-16 ***
## mileage     -0.06491    0.02104  -3.086  0.00213 **
## IndLoc       1.38544    0.61272   2.261  0.02411 *
## age         -1.24415    0.19656  -6.330 4.83e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.346 on 596 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.3451
## F-statistic: 106.2 on 3 and 596 DF,  p-value: < 2.2e-16
```

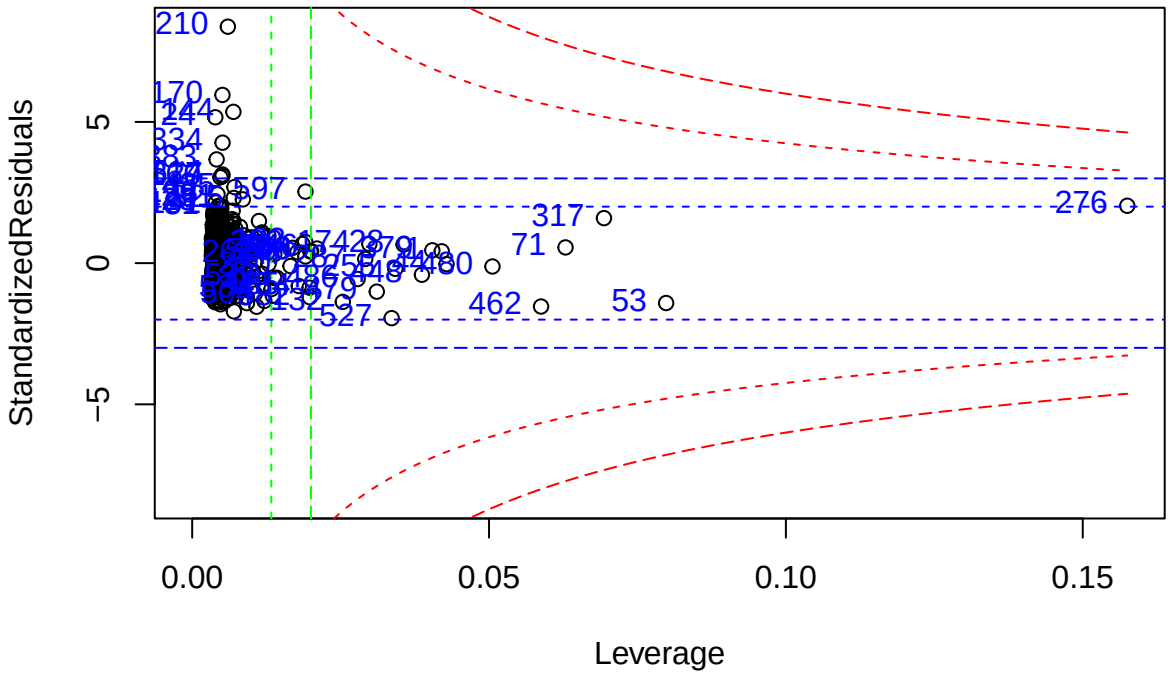
Comments

After playing around with different models, the final model I have come up with is $\hat{\text{price}} = 31.53469 - 0.06491 * \text{mileage} - 1.24415 * \text{age} + 1.38544 * \text{IndLoc}$. I decided to go with this combination of variables because they produced the highest r^2 of 0.3484 and adjusted r^2 of 0.3451 while still maintaining p-values that were significant enough to reject the null and in favor of the alternative.

Assess

Figures:


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ log(mileage) + sqrt(age^2) + IndLoc)
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(price ~ log(mileage) + sqrt(age^2) + IndLoc)
210
170
144




```
##
## Call:
## lm(formula = price ~ log(mileage) + sqrt(age^2) + IndLoc, data = car_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.899  -4.990  -1.951   3.257  57.482
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.5435     1.4229  26.386  < 2e-16 ***
## log(mileage)  -2.4941     0.5060  -4.929 1.07e-06 ***
## sqrt(age^2)   -1.2864     0.1394  -9.231  < 2e-16 ***
## IndLoc         1.4942     0.6040   2.474   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.274 on 596 degrees of freedom
## Multiple R-squared:  0.3611, Adjusted R-squared:  0.3579
## F-statistic: 112.3 on 3 and 596 DF,  p-value: < 2.2e-16
```

Comments

Following the addition of a log transformation on mileage and a sqrt transformation on age, we are able to address the LNE conditions for this model. Additionally, I did some data cleaning to help remove some outliers in hopes of creating some better plots to meet the LNE conditions. As we can see, in our residuals vs. fitted plot, we have minimal flaring and a equal amount of points above and below our axis telling us that we have met our linearity and equal variance conditions. Additionally, our Q-Q plot looks pretty good, and our cooks plot has minimal values with residuals greater than 3 and leverages greater than 0.15 which gives us fewer points with greater influence on our model.

Use

For a car in New York that is 3 years old and has 40,000 miles on it, the prediction interval is [8.94063, 38.77526]. This means that I am 95% confident that the true price of a car in Ohio that is 3 years old with 40,000 miles on it falls within the interval of $8,940.63 and $38,775.26.

Conclusion

Based on the analysis performed in this project, it was found that there is a difference in prices between New York and Ohio for used Jeep Wranglers after accounting for factors such as mileage and age. While the difference in prices between the two location was not very large, where Ohio had the slightly higher average prices, the analysis revealed that mileage and age were the most significant predictors of a car's price. In my opinion mileage had a stronger correlation over age because it was demonstrated to have a greater correlation coefficient. The model developed for predicting car prices explained about 35% of the variability in price. Overall, while the model provides some insights into the factors that influence a Jeep Wrangler's price, additional information such as vehicle condition and market demand could help better predict a Jeep Wrangler's price as well.