Jenna Beauchain
Juan Betancor
Myles Nelson-Atkins
Max Nussbaum
Tetsuro Furuhashi
April 29th, 2024
Math 123
Professor Ream

Introduction

Our research question explores whether we can successfully build a model that predicts the price of a used car given its year, odometer, and condition. Newer models are valued higher as a younger car will benefit from technological and safety advancements. Mileage affects the car's price negatively with a higher mileage associated with a higher likelihood of encountering performance issues due to wear and tear. A car's condition describes both the looks and usability of the car. We expect a better condition to indicate a higher price. We focus on how these factors may influence the pricing of the cars. Our analysis involves using regression plots to determine the nature of these relationships. We expect that each factor will play a significant role independently and collectively in determining the final price of a used car.

The regression model equation is
Y = alpha1 + beta1 x year + beta2 x odometer + beta3 x condition
intercept (alpha1): This intercept represents the estimated value of the price when the year, odometer, and condition are equal to zero. In our model, this represents the estimated price of a car with zero years, mileage, and a baseline condition.
Year (beta1): This coefficient represents the change of the price of the car for each year that passes. In our model, this indicates how much the price of the car increases or decreases for each year added.
Odometer (beta2): This coefficient represents the change in price of the car for every 1 mile increased. In our model, this shows the impact of mileage on the price of the car.
Condition (beta3): This coefficient represents the change in price of the car for for each condition category, compared to the baseline condition. In our model, it indicates how different conditions affect the price of the car compared to the baseline condition.

The hypotheses were consistent across each condition with the null being there is no effect of the independent variable on the dependent and the alternate being there is an effect of the independent variable on the dependent. In context to this regression model, if the null hypothesis is not rejected the coefficient will be equal to zero, and if the null is rejected the coefficient will not equal zero meaning it has an effect on the dependent variable.

These six graphs demonstrate the relationship between three aspects of used cars: year, mileage, and condition. Our data consisted of 192,783 different used car listings on Craigslist. In the data we used, there were initially more columns of data which we removed and "cleaned up," only taking cars with a price range between $1,000 and $57,990. We altered the data to only include the cars with years between 2000 and 2020 and odometer readings from 10,000 to 285,000 which averages to around 94,723 miles.

The outcome of our data exhibited a small p-value throughout all of our coefficients except for the "new" condition which had a slightly larger p-value of 0.085. Since almost all of the p-values are about 0.05 or lower, our data is statistically significant in predicting the price of the car.

1) **Summary of the Data**
    a) Size: the size of our data is 192,783 different listings on Craigslist
    b) We had more datapoints, but we "cleaned" it up, only taking cars with prices ranging from $1,000 to $57,990
    c) The years were also cleaned up to only include cars with years between 2000 and 2020.
    d) The odometer readings range from 10,000 to 285,000 averaging around 94,723 miles.
    e) There is the variable "state" that we did not choose to analyze as we did not think it would be significant towards the price of the car.
2) **Model Coefficients**
    a) **Intercept**. Our intercept was -1,942,000. This is incredibly unrealistic as this would be the price of a car never driven, in excellent condition, from the year 0.
    b) **Year**: With each additional year, the car's value increases by approximately $976.48 with a standard error of $5.277. This aligns with our expectations because the bigger the year, the newer the car.
    c) **Odometer**: With each additional mile on the odometer, the price increases by about $0.064 with a standard error of approximately $0.00045. This aligns with our expectations because as the mileage on the odometer increases, so does the wear and tear on the car.
    d) Condition: All of this is compared to a baseline condition of "Excellent".
        i) **Salvage**. A car given a salvage condition reduces the price for that car by $5,123.90 with a standard error of $476.048. This reflects the low value of the car given its poor condition.
        ii) **Fair.** A condition of fair usually means a price drop of roughly $1,007.86 with a standard error of $148.104.

iii) **Good**. A condition of new increases the price by $2060.65 with a standard error of $41.410. This is inconsistent with our expectations that a good car should be worth less than an excellent one. We think this could be due to a non-standardized grading level. A car considered to be in a "good" condition by one person may be valued as high as an "excellent condition" by another. The ambiguity of the rating "good" could be at fault for the unexpected coefficient. This could also be due to the number of cars in good condition as nearly half of our dataset were cars in "good" condition.

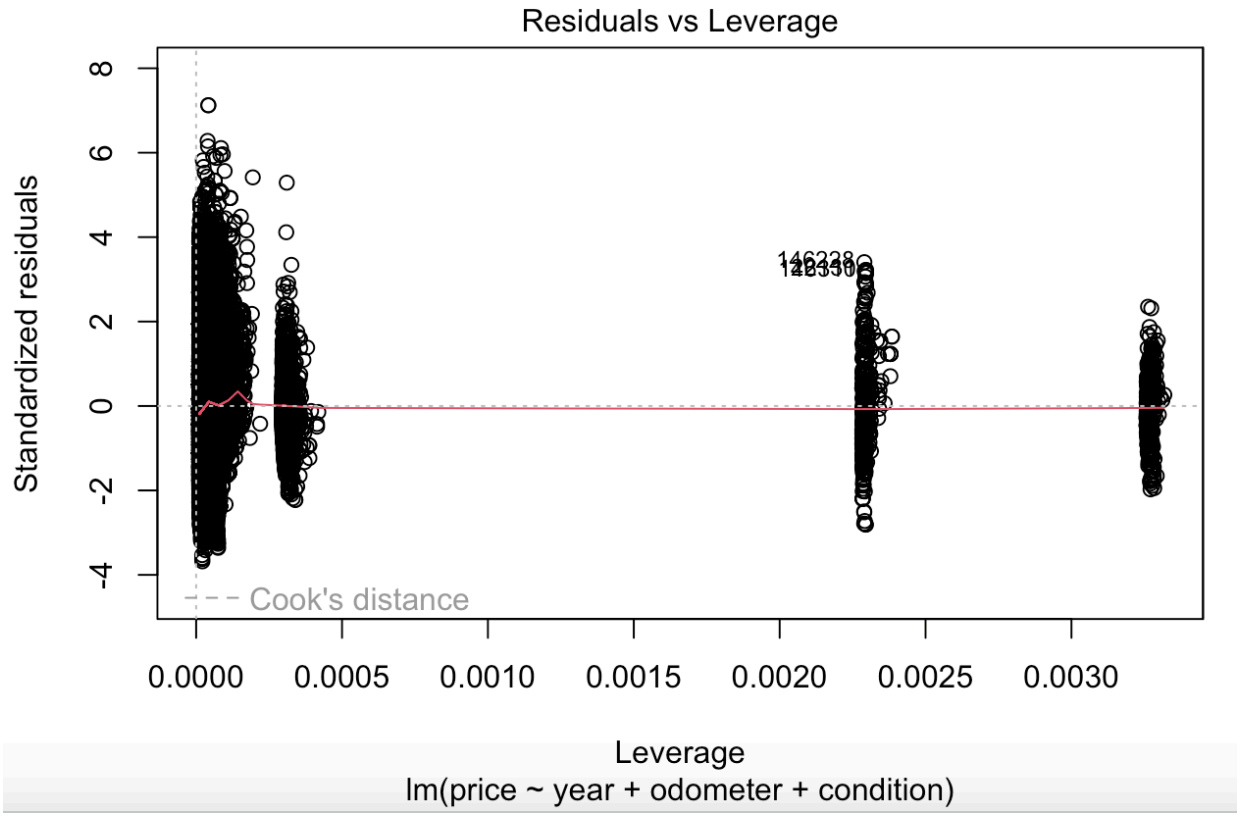
iv) **Like New.** A car given a like new condition will see a decrease of $194.62 with a standard error of $72.578. This aligns with our expectations.

v) **New**. A car given a new condition sees a decrease of $687.41 with a standard error of $398.83. This doesn't align with our expectations as we would expect it to be a lot less than Like New. We think this may be because of the platform our data is coming from. For a car to be on Craigslist and considered "New", it would be an undesirable car. This could also be due to an extremely low amount of "new" cars as only 438 of the cars were listed as "new".

3) Results and quality of the model

a) **R-squared**: 0.504. This indicates that roughly 50.4% of the variability in prices is explained by the model. While this is not too bad (our original attempts yielded a 13% R-squared variable), this does indicate that there are factors that are not considered in this model. These factors may include the make/model of the car and the time when the car was listed.

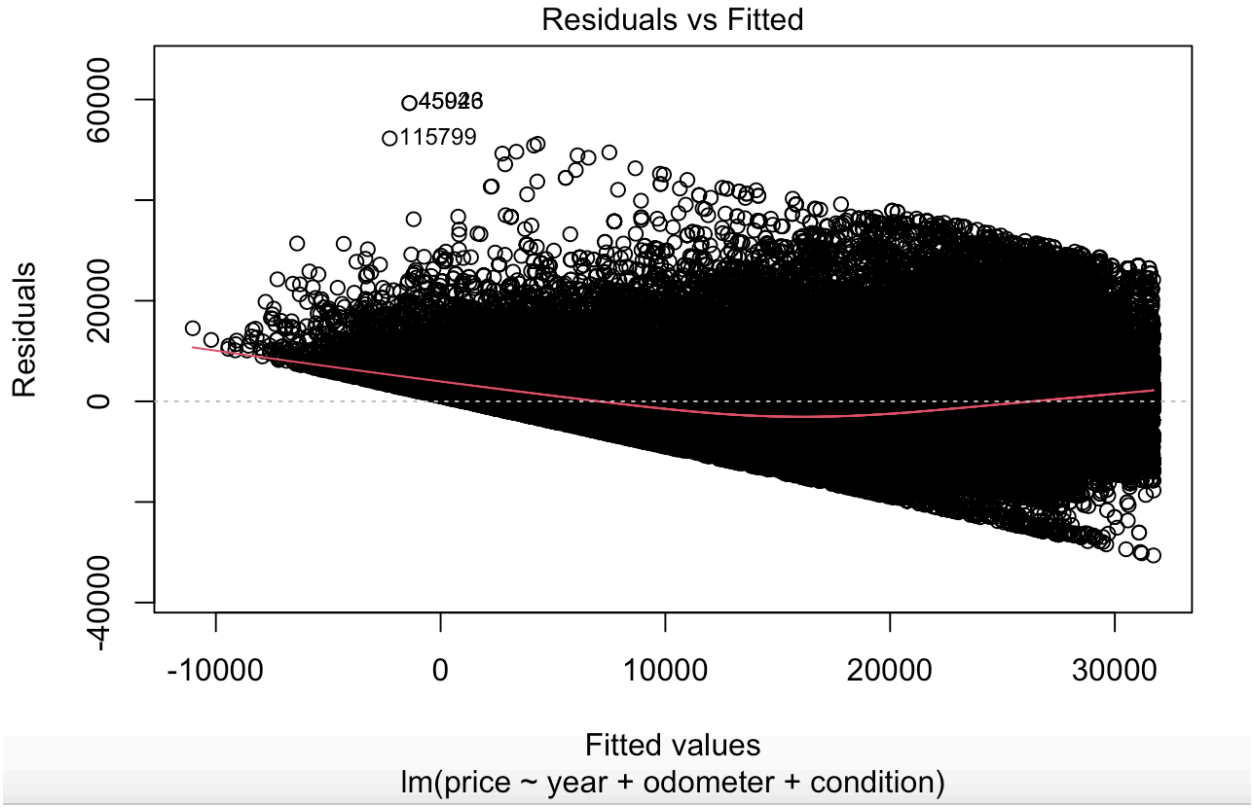
b) **Residuals**: We had a residual standard error of 8323. An improvement of 8.64 million on previous models. The model also has 192775 degrees of freedom



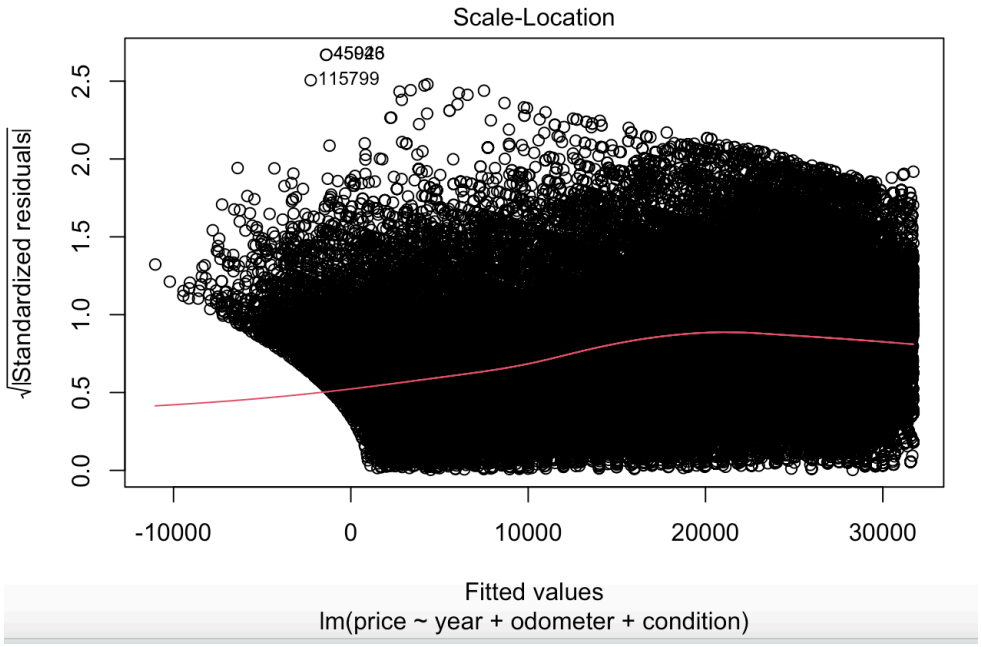
This graph shows a Residuals vs Leverage plot with Cook's distance in dashed lines. We use the year created, odometer, and condition and see its relation to the cars' prices. The vertical axis shows the standardized residuals from the linear regression model. The horizontal axis shows the influence an individual point has on the fitted model. Cook's distance is a measure of how much an observation affects the regression model.

A good number of the points are close to the horizontal zero line, while we can see some outliers far above and below. This is not exactly ideal for a fitted model. Similarly for leverage, most points are close to zero, but we have two groups of points between 0.0020 and 0.0025, and above 0.0030, which is noticeably higher than the majority. The points with both high leverage and residuals are concerning, since they not only have extreme X values but also deviate significantly from the predicted values, potentially messing with the regression results. We can say that these points are influential points in the model.

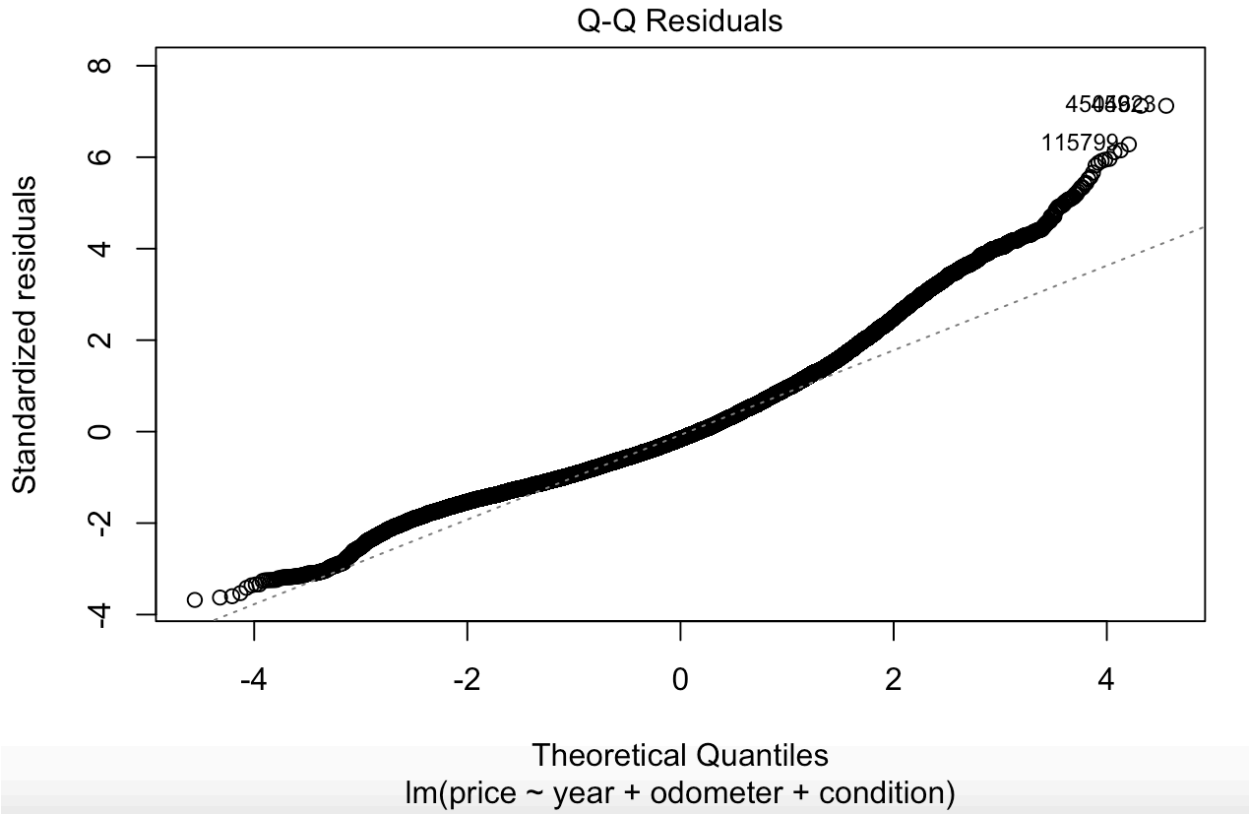
In summary, this plot shows that the model does not rely only on a few data points, and provides reliable and generalizable insights. It could be useful to revisit the outliers to check if they are error inputs or not. If they are valid data points, we can use more powerful regression techniques to lessen the influence of those points for a more stable model estimation.



This graph shows a a linear regression model on a Residual vs. Fitted plot that uses the year created, odometer, and condition to predict the car's price. This plot examines the differences between observed and predicted values, using them to find potential issues with the regression model. The mean of the residuals in this graph is represented by the red line. This line represents the "ideal" model. Instead of scattered around zero, our residuals show a pattern where variability increases along with the fitted values. The widening shape of the graph indicates increasing variance in residuals as the predicted price increases



The Scale-location plot helps visualize how equally spread the residuals are along the range of predictors. Ideally the residuals would be randomly distributed with a constant variance across the fitted values range. However in this plot the residual values do not have a constant variance. The scale-location plot helps access the assumption of homoscedasticity and because of the differing variance of residuals, this plot indicates that said assumption could have potential errors. This could lead to unreliable statistical tests and confidence intervals. In future models this issue of homoscedasticity could be solved by making transformations of the dependent variable in the model such as logging the price. This could potentially fix the error in the variance.



This Q-Q residuals shows the standardized residuals from a linear regression model, which predicts price based on variables such as year, odometer, and condition.
Since the points are supposed to follow the line, and the graph shows heavy tails, implying potential outliers or heteroscedasticity, where residual variance increases with the response variable's value. This deviation from normality can affect the validity of inferential statistics in regression, such as t-tests. To fix this, we could use robust regression methods to improve the model's assumptions and accuracy.

Conclusion

While our model has its shortcomings, there are indications of statistical significance. We believe that, specifically with used cars, there are too many variables to fully develop a 100% accurate model. Variables such as the car's history, color, make, and model all effect the price in ways that are extremely hard to predict. Overall, the R-Squared value of 50.4% shows that, while variability exists, our model can correctly explain over half of the variability in car prices. The variability the outcome showed suggests that there are other factors we did not account for in our model. In the future, we would look to include more variables like the make/model of the car, geographical location, or economic conditions to create a more accurate outcome.

Statistical tests: When performing regression analysis, statistical tests are conducted to assess the significance of each coefficient in the model. The most common test is the t-test, which determines whether the coefficient for each independent variable is significantly different from zero.

Null hypothesis (H_0): The null hypothesis states that there is no effect of the independent variable on the dependent variable. In the context of regression, the null hypothesis for each coefficient (beta) is that the coefficient is equal to zero, indicating no effect of that independent variable on the dependent variable.

Alternative hypothesis (H_1): The alternative hypothesis states that there is an effect of the independent variable on the dependent variable. In the context of regression, the alternative hypothesis for each coefficient (beta) is that the coefficient is not equal to zero, indicating that the independent variable has a significant effect on the dependent variable.

When reporting p-values, it's important to include whether the null hypothesis is rejected or not for each coefficient. A low p-value (typically less than 0.05) indicates that the coefficient is statistically significant, and the null hypothesis can be rejected in favor of the alternative hypothesis. Conversely, a high p-value suggests that the coefficient is not statistically significant, and there is insufficient evidence to reject the null hypothesis.

For example, if the p-value for the coefficient of the year variable is 0.02, you would report something like:

"The p-value for the coefficient of the year variable (

$\beta 1 \beta$

1

) is 0.02. Since this p-value is less than 0.05, we reject the null hypothesis and conclude that the year variable has a statistically significant effect on the price of the car."

Including this information provides clarity on the significance of each coefficient in your regression model and helps readers understand the reliability of your findings.

```
Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)       -1.942e+06  1.065e+04 -182.384  < 2e-16 ***
year               9.765e+02  5.277e+00  185.027  < 2e-16 ***
odometer          -6.431e-02  4.514e-04 -142.474  < 2e-16 ***
conditionfair     -1.008e+03  1.481e+02   -6.805 1.01e-11 ***
conditiongood      2.061e+03  4.141e+01   49.762  < 2e-16 ***
conditionlike new -1.946e+02  7.258e+01   -2.682  0.00733 **
conditionnew      -6.874e+02  3.988e+02   -1.724  0.08479 .
conditionsalvage  -5.124e+03  4.760e+02  -10.763  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```