Abstract

This study addresses the interaction between various factors and their impact on the pricing of Ford models in the UK, focusing on fuel type, transmission type, and the combined effects of mileage and age. Using various statistical methods and hypothesis testing, I seek to uncover statistical significances that influence vehicle prices.

The first research question explores the relationship between fuel type (diesel, petrol, and other which includes electric and hybrid) and road tax on the pricing of Ford models. I hypothesized that vehicles with higher road taxes, particularly electric and hybrid models, command higher prices due to their lesser environmental impact despite their lower road tax rates. Preliminary models show a significant interaction between road tax and fuel type, suggesting that cleaner vehicles with higher taxes tend to be more expensive.

The second question examines how transmission type (manual, automatic, and semi-automatic) and fuel efficiency (measured by miles per gallon) affect used Ford car prices. I predicted that cars with better fuel efficiency are generally cheaper unless they feature semi-automatic transmissions, typically found in high-end sports cars, which might invert this trend. Results indicate significant interactions, suggesting that the inclusion of transmission type improves the model than.

The third study investigates the impact of mileage and age on Ford vehicle pricing, employing a Weighted Least Squares regression model to accommodate potential heteroscedasticity. The findings suggest that both higher mileage and older age significantly reduce vehicle prices, confirming the intuitive market logic that older and more used vehicles are cheaper.

Each section of the study involves rigorous diagnostics, including tests for normality, heteroscedasticity, and influential points, leading to several model adjustments such as robust regression and transformations like the Box-Cox to ensure the validity of the regression results. The study concludes that while mileage, age, fuel type, and transmission type significantly influence car prices, the predictive power of the models varies, reflecting complexities in vehicle pricing dynamics influenced by regulatory, environmental, and technological factors.

Background

The automotive industry is a complex and ever-evolving market where vehicle prices are influenced by a myriad of factors including technological advancements, regulatory changes, and consumer preferences. As environmental concerns drive stricter regulations, vehicles that utilize cleaner technologies such as electric and hybrid models are becoming increasingly prominent. These vehicles often benefit from lower road taxes due to favorable policies aiming to reduce carbon emissions. However, the actual impact of these policies on vehicle prices, particularly when combined with other factors like fuel type, remains poorly understood.

Similarly, transmission type has been a key factor in vehicle pricing, with shifts in consumer preference towards automatic and semi-automatic transmissions reflecting changes in technology and driving habits. The relationship between transmission type, fuel efficiency, and vehicle pricing is intricate, as newer, more fuel-efficient vehicles tend to be more expensive, potentially offsetting the cost benefits of improved mileage.

Furthermore, traditional factors like mileage and age continue to play critical roles in determining used vehicle pricing. Higher mileage typically indicates more wear and tear, which reduces vehicle prices. Similarly, older vehicles tend to depreciate in value over time. Understanding the combined effect of mileage and age on pricing can help consumers make better purchasing decisions and allow businesses to better position their products in a competitive market.

Given these dynamics, this research aims to dissect the interaction between these variables—fuel type, road tax, transmission type, mileage, and age—and their collective impact on the pricing of Ford models in the UK. This study uses statistical models to analyze data from various listings of Ford vehicles, applying robust regression techniques to account for outliers and potential biases in the data. By exploring these relationships, the study seeks to provide insights that could influence both consumer choice and policy-making in the automotive sector.

Literature Review:

Angadi, P. (2023). Data Analysis Project on Used Car Sale Price. ResearchGate. https://www.researchgate.net/publication/370523018_data_analysis_project_used_car_sale_price

Comparing the "Data Analysis Project on Used Car Sale Price" by Priyadarshini Angadi and our study, there are notable similarities and distinctions in methodologies and findings.

Both studies aim to find the determinants of used car pricing, with Angadi focusing on the Toronto market and our study centered on UK's Ford models. Angadi's research evaluates a broader range of control variables, such as make, body type, and engine, while our study delves into the interaction effects between variables such as fuel type, transmission, mileage, and age.

In terms of methodology, Angadi employs a logarithmic transformation for price and mileage to facilitate interpretation, similar to our approach that includes transformations and weighted least squares regression to address heteroscedasticity. This transformation is crucial as it adjusts for non-linearity and variance inconsistency, which are prevalent issues in regression analyses of economic data.

Both studies acknowledge the significant impact of mileage and vehicle age on pricing, confirming the intuitive market logic that higher mileage and older vehicles tend to be less expensive. Angadi's results are similar with our findings, as she also identifies a negative correlation between mileage and price and a positive correlation with the vehicle's year of manufacture. These insights align with the common understanding that newer and less-traveled vehicles fetch higher prices in the used car market.

One aspect where our study stands out is the application of diagnostic checks and model adjustments, including robust regression techniques to address outliers and potential biases in the data, which are not detailed in Angadi's study. This rigorous approach to ensuring the validity of our regression results enhances the reliability of our findings.

In conclusion, while both studies contribute valuable insights into used car pricing, our STAT 512 report extends the analysis by incorporating interactive terms and robust regression techniques, offering a better understanding of the UK's used car market. The findings from both studies can inform buyers, sellers, and policymakers, improving market transparency and efficiency.

Research Question #1

This research question investigates whether there is a statistically significant interaction between the fuel type of the Ford model and the amount of road tax that is required for that Ford model on its selling price. Generally speaking, the higher the road tax the higher the price for the car. However, electric and hybrid vehicles, being clean-energy vehicles, will likely pay less road tax when compared to petrol and diesel cars. There are additional factors that go into the UK road tax, such as list price and vehicle type. Perhaps, the price changes more significantly with electric cars that have higher road tax. Perhaps, it is the reverse effect. Nonetheless, this question is worth investigating to understand the nature of the modern day car market. What is certainly known is that, generally speaking, more expensive cars pay higher road tax. It is a matter of how much more tax that is the question.

As defined by our model, X_3 represents quantitative variable road tax (pound sterling) and X_8 and X_9 represent dummy variables for the fuel types "diesel" and "other" respectively. The baseline level is "petrol" and the "other" category represents all fuel types other than petrol and diesel. This category essentially includes all electric and hybrid Ford models. Y, of course, represents the price in pounds sterling of the Ford model. Generally, we expect that models in the "other" category will have a lower road tax compared to petrol and diesel cars.

Given this information, we can represent a full and reduced model as such:

$$Y = \beta_0 + \beta_3 X_3 + \beta_8 X_8 + \beta_9 X_9 + \beta_{38} X_3 X_8 + \beta_{39} X_3 X_9 + \varepsilon$$

$$Y = \beta_0 + \beta_3 X_3 + \beta_8 X_8 + \beta_9 X_9 + \varepsilon$$

The reduced model essentially excludes the interaction terms between fuel type and road tax. With these two models, we can define our null and alternative hypothesis as the following:

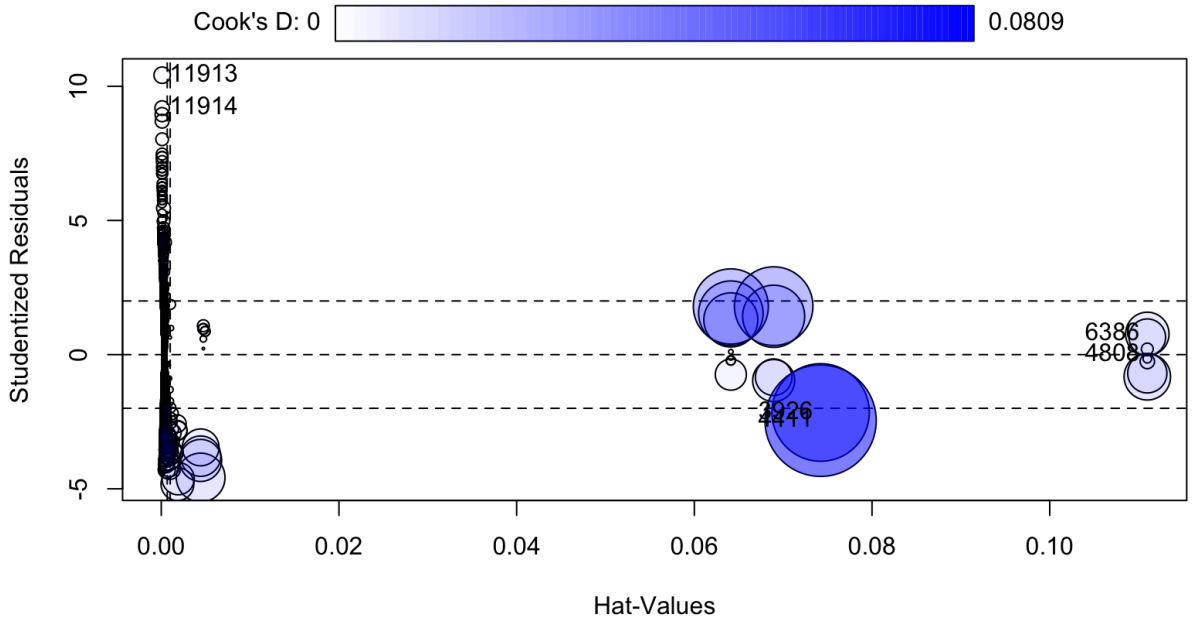
$$H_0: \beta_{38} = \beta_{39} = 0$$

$$H_A: \beta_{38} \neq 0 \wedge \beta_{39} \neq 0$$

The initial full model has a R-square of 0.2498 and adjusted R-square of 0.2496. This indicates that there is still a lot of unexplained variance in the price data that is not explained by the predictor variables.

Diagnostics & Remedy

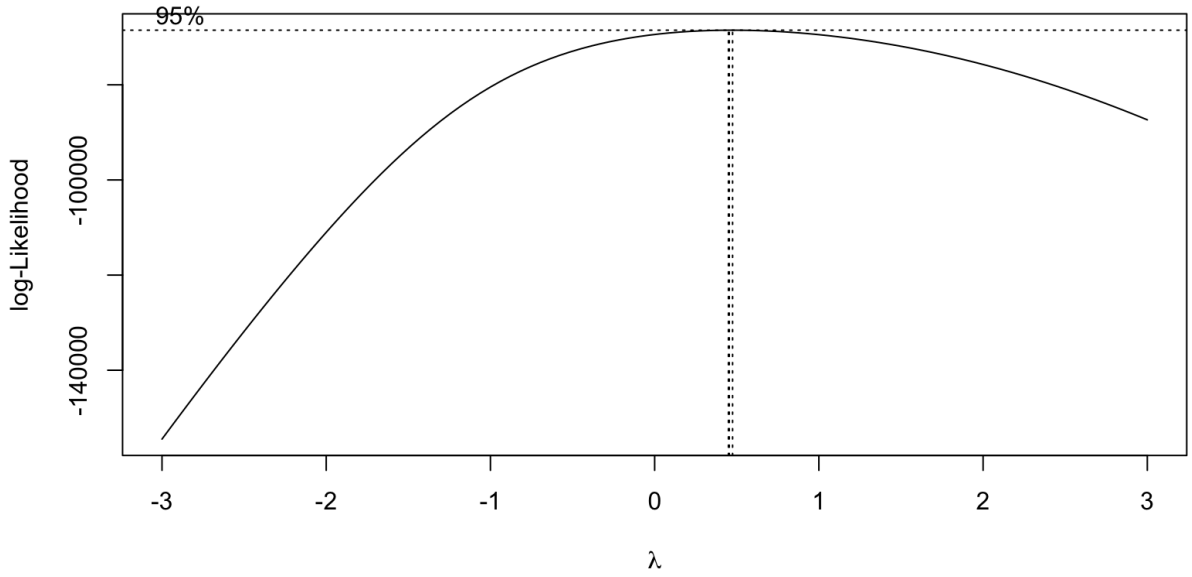
First, I will look at outliers and influential points in the data and see if there is a way to remedy it using a robust regression. My criteria will be based on the Studentized deleted residual, Hat value, and Cook's distance.



The points indicate significant influential points in a few data points based on these three values combined. Based on this, I will try to use a robust regression to reduce the impact of these outliers. Unfortunately, the robust regression still results in influential points based on the studentized deleted residual value.

Next, I will test the residuals of my model to check for non-normality and heteroscedasticity. I will test non-normality using the Shapiro test on the residuals of my model. From this, I get a W value of 0.94054 which results in a p-value close to 0. This indicates that the plot of the residuals don't follow a normal distribution by a statistically significant amount. The Breusch-Pagan test gives a value of 675.28 which results in a p-value close to 0, indicating the heteroscedasticity is present among the residuals.

To remedy this issue, I first used the Box Cox transformation on Y. The optimal lambda based on the MLE is 0.4545 as shown by the plot below:



Thus, the optimal transformation on Y with the highest MLE is:

$$Y_{new} = Y^{0.4545}$$

Unfortunately, after this model was applied to the data, it neither solved the heteroscedasticity issue nor did it remedy the non-normality of the residuals. The Breusch-Pagan test still showed a high value of 862.58 and the Shapiro test with a value of 0.9677. Both these values result in p-values close to 0. Furthermore, the new model shows almost no-improvement on the R-square value.

As such, it is clearly ill-advised to follow through with this approach.

The next best approach would be to follow through with a weighted least squares regression model (WLS) to remedy the heteroscedasticity. After adjusting the weights and implementing the model, the new R-square value is 0.3568 and adjusted R-square is 0.3566. This marks a significant improvement from the initial full model as it had an R-squared and adjusted R-square 0.10 lower. The Breusch-Pagan test on the WLS model shows a value of

0.00020014 which results in a p-value close to 1. This indicates that the residuals have virtually no heteroscedasticity in the WLS model. Given these improvements, we will continue using this model to test the hypothesis.

The test statistic (F-value) was 304.35 which is based on the difference in SSR from the full and reduced model divided by 2 (difference in degrees of freedom) divided by the MSE of the full model. The F-statistic results in a p-value close to 0 indicating that the interaction between the fuel type and the road tax amount on the price of the Ford model is statistically significant. This makes sense, electric cars should expect to pay a higher price premium on higher road tax since the road tax on those types of vehicles are already very small. For electric cars to pay higher road tax, it would have to mean that they are more expensive rather than just being more carbon neutral.

The equation of the model is:

$$\text{Price} = 8204.9329 + 28.5857*\text{tax} + 592.5001*\text{diesel} + 4670.2320*\text{other} + 19.4737*\text{tax}*\text{diesel} + 68.2862*\text{tax}*\text{other}$$

The 10 fold-cross validation results in a RMSE of 4106.679 and R-squared of 0.2508439. These results are somewhat decent given the variables being used for the model.

The potential issues with the predictive power of the model mostly lie in the data. Although it makes clear what the car's fuel type is, the road tax that a car pays can be more ambiguous. For example, how does the data account for large pickup trucks compared to small electric cars? Also, what other factors are accounted for via road tax which does not just include carbon emissions? The tax laws around these cars could be a bit ambiguous making it hard to exactly predict the price of a car solely based on road tax and fuel type. Given that some issues have been found with the data, there could be other potential misinputs, especially surrounding the road tax values. Then there are also practical concerns regarding the usability of this model. Does a potential buyer have access to road tax numbers on vehicles? Even if they do, it wouldn't account for a significant portion of the price variability.

Research Question #2

The second research question aims to test the significance of transmission type and its interaction terms on a model containing mpg to predict the price of a used Ford car. Generally, vehicles with better mpg tend to be more economical and cost less than better cars with worse mpg. This is because higher-end vehicles tend to have bigger engines to reach higher speeds or support their bigger size. The bigger engine size thus consumes more fuel than their cheaper counterparts with a more economical engine size. So vehicles with worse mpg could have a higher price while better mpg could have a lower price. There is also the possibility of highly used vehicles having worse mpg from years of use. Thus we predict the extreme ends of the price spectrum to have worse mpg. We would like to see if transmission type would improve the model to predict the price of the car as there is no direct link between the mpg and transmission type. Vehicles with manual transmissions are probably older than vehicles with automatic transmissions because automatic vehicles have dominated the market than their counterpart in recent years. With this information, automatic vehicles will probably cost more than manual vehicles. Semi-automatic cars, similar to manual cars, are less common nowadays but they made for high-end sports cars. With this information, semi-automatic vehicles will probably have a high cost for the lower mpg but lower cost as the mpg increases because they probably won't be popular among manual and automatic transmission car enjoyers.

With the research question in mind, we now have to create a hypothesis test along with a reduced and full model. The variables we will be using to predict the price of the car (Y) are MPG (X4) and transmission type. MPG is a quantitative variable while transmission type is a qualitative variable. Since the transmission type is qualitative,

it is important to split it up into a baseline with variables for the other transmission types. The baseline we will be using is manual transmission while automatic transmission (X6) and semi-auto transmission (X7) will be the variables.

With the background covered, we can now provide the full and reduced models:

$$Y = \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 + \beta_{46} X_4 X_6 + \beta_{47} X_4 X_7 + \varepsilon$$

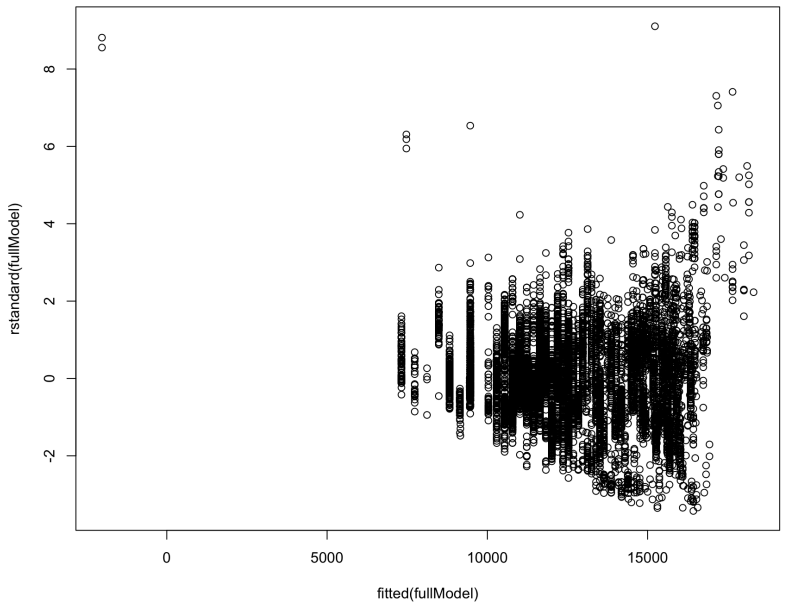
$$Y = \beta_4 X_4 + \varepsilon$$

The reduced model excludes the transmission types and the interaction terms making it solely based on mpg. This is to test whether or not transmission type and its interaction terms are significant or not. With the models, we can also define the hypotheses:

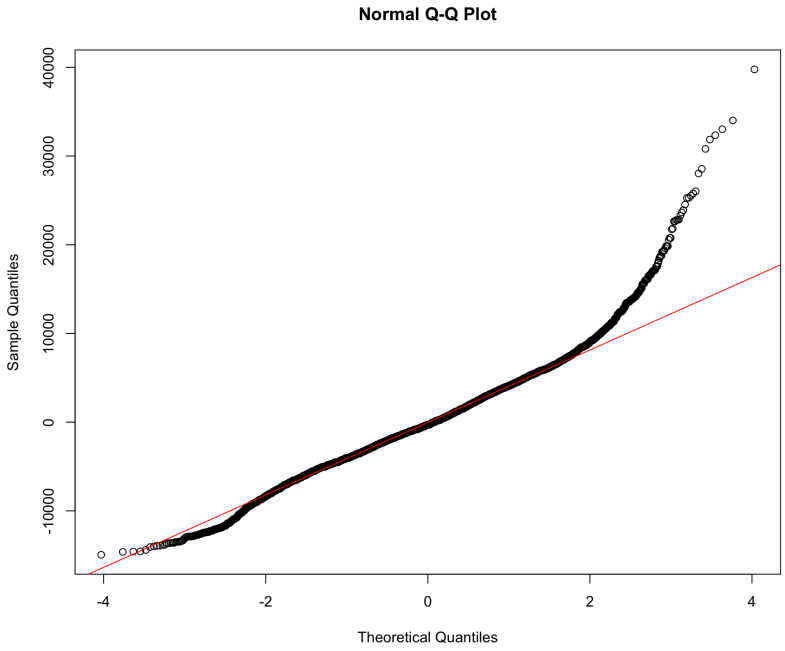
$$H_0: \beta_6 = \beta_7 = \beta_{46} = \beta_{47} = 0$$

$$H_a: \beta_6 \neq 0 \; or \; \beta_7 \neq 0 \; or \; \beta_{46} \neq 0 \; or \; \beta_{47} \neq 0$$

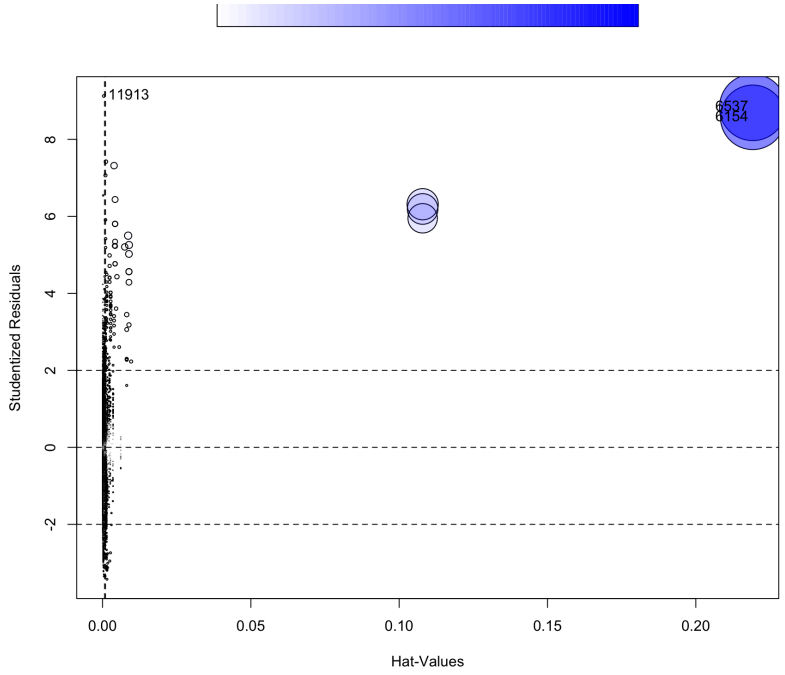
With the hypotheses and models defined, we can test the model for heteroscedascity, normality, and outliers. There isn't any need to check for multicollinearity because we are only working with one quantitative variable. To begin, we checked for heteroscedascity using a Breusch-Pagan test and a regression plot. The Breusch-Pagan test yielded a p-value < 2.2e-16 signifying non-constant variance. The regression plot was similar, showing non-constant variance.



Next was to check for normality. Using the normalized Q-Q plot, we can determine if normality was an issue by observing how close to the line the points fit. Additionally, we also ran a Shapiro test randomly sampling the dataset to check for normality. From the plot below, there does seem to be a normality problem due to the right tail deviating heavily from the line. The left tail deviates slightly but not as heavily as the right tail. The Shapiro test yielded a p-value < 2.2e-16 supporting non-normality.



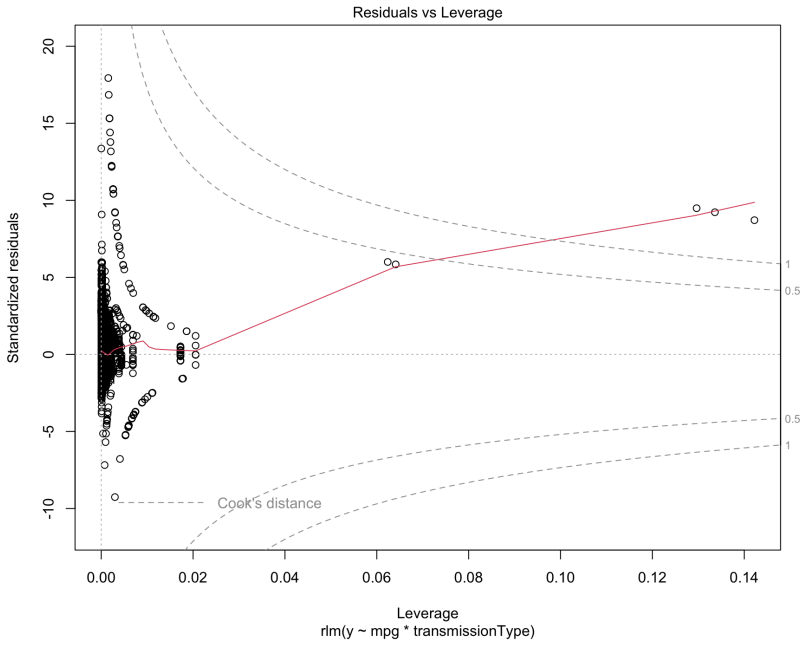
The final check is for outliers and points of influence. We used an influence plot to determine points of influence. The plot below shows areas of significance where the darker the shade of blue an area has, the more influential the point is. The right side has a few points of major influence while the middle has points of influence too.



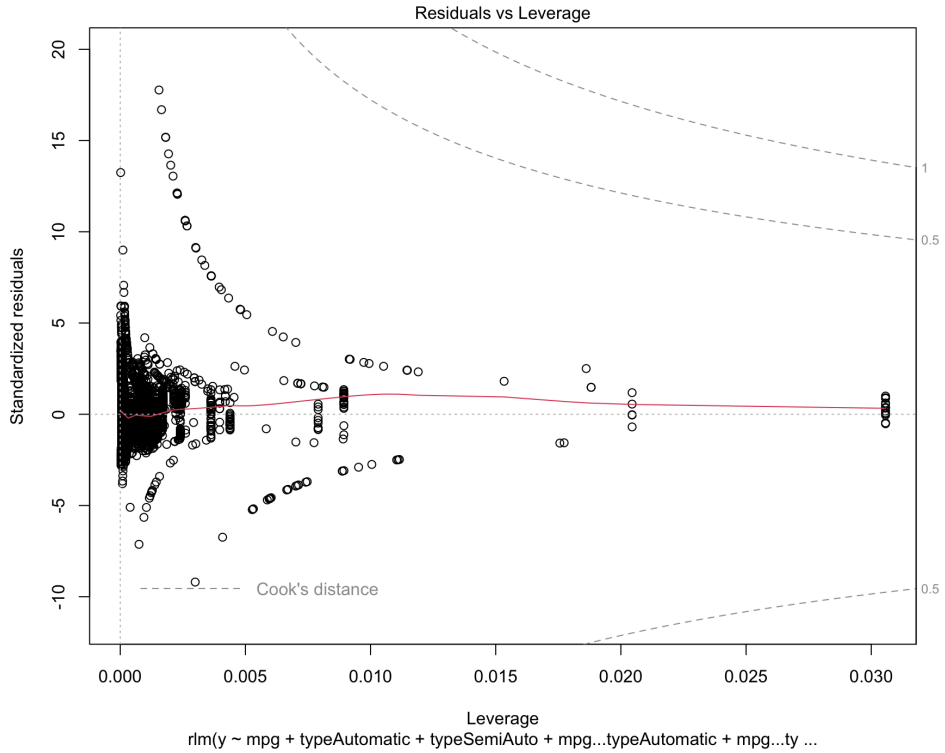
So from these tests, we have deemed the model to have an issue with variance, normality, and outliers. To remedy this we will try the box-cox procedure for normality and variance, weighted least squares regression for variance if the box-cox doesn't work, and a robust regression for the outliers. For the box-cox procedure, we got a

lambda of 0.5 and made a new model using the lambda to adjust the Y value. After making the model, we check for both variance and normality with the Breusch-Pagan and Shapiro tests and their respective plots. The Breusch-Pagan test and Shapiro tests yield a p-value < 2.2e-16 indicating non-constant variance and non-normality. So the box-cox procedure fails, making the WLS regression necessary.

For the WLS regression, we make the model using the full model instead of the box-cox model because the box-cox model failed. After making the WLS model, the Breusch-Pagan test yielded a p-value of 1 meaning the variance is finally constant. The Shapiro test still had a significant p-value meaning the normality issue still persists. To address the outliers, we used a robust regression model from the WLS model. From the robust model, the plot of interest is the one focused on outliers. The plot below shows some points of major influence on the right even after robust regression. It is best to remove those points when robust regression isn't able to reduce their impact.



After removing those points, we made another robust model with new adjusted weights to accommodate the changes. The new plot of residuals vs leverage has a much better distribution of points which can be seen below. While some points still seem to have minor influence, there are no more points of major influence.



Again, we check for variance and normality. Using the same tests as before, we pass the test for variance and fail the test for normality. The normality issue seems to be persistent, but we have solved variance and influential points. Now to test our hypothesis question, we can use the summary of the adjusted robust model to determine the significance of the variables. The critical t-value is t(1-0.05/2, 17955) = 1.960096. The summary of the robust model provides t-values for each variable, if the absolute value of their t-value is greater than the critical t-value then it is significantly different from 0. Additionally, we can also make confidence intervals and if the interval contains 0 then it is not statistically different from 0. Using both, we deem automatic transmission type to not be significant but all other variables to be significant. This means that we reject the null hypothesis and accept the alternative.

The final equation of the model:

$$Y = 17788.8998 - 105.0861X_4 - 1077.2596X_6 + 4048.1617X_7 + 77.3608X_4X_6 - 39.1338X_4X_7$$

Where Y is the price, X_4 is the MPG, X_6 is automatic, and X_7 is semi-auto.

A major issue with the model is that it doesn't have a normal distribution. Our first remedy, the box-cox procedure, didn't solve the issue meaning that the model might not be quadratic in nature but exponential. Another attempt to address this was with robust regression but that also didn't solve the problem. From this, the best model we could choose was one that addressed the other issues of heteroscedascity and outliers. By using WLS, robust regression, and removing points of extreme influence, our model addressed those two issues. From the model, while automatic transmission type was deemed not different than 0, the interaction term was. This means that for lower, mpg diesel and automatic vehicles have similar prices but as the mpg increases, the price for diesel vehicles decreased more than automatic ones. Further analysis on the age of the vehicle could give further insight as to why prices are similar for the two transmission types when mpg is low. As predicted, semi-automatic transmission did have a higher initial cost for low mpg, but the price then dropped faster than diesel transmission as mpg increased.

Research Question #3

Introduction

The dynamic landscape of the automotive industry is driven by various factors that influence vehicle pricing. This report focuses on the relationship between two key variables: mileage and age, and their impact on the pricing of Ford vehicles. The objective is to determine if mileage and age are statistically significant predictors of vehicle price and how they interact to influence pricing decisions.

Methodology

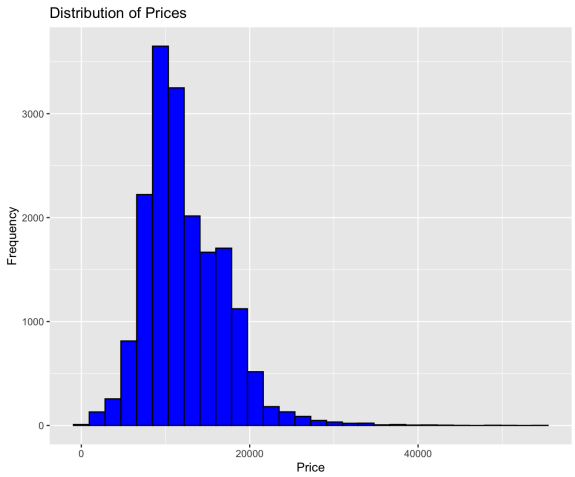
The data comprises various listings of Ford vehicles, including details such as price, mileage, and age. The analysis begins with data cleaning to ensure accuracy and relevance, followed by an exploratory data analysis (EDA) to understand the distribution and relationship between the variables. To analyze the effects of mileage and age on vehicle price, a Weighted Least Squares (WLS) regression model was employed. This approach was chosen to account for potential heteroscedasticity.

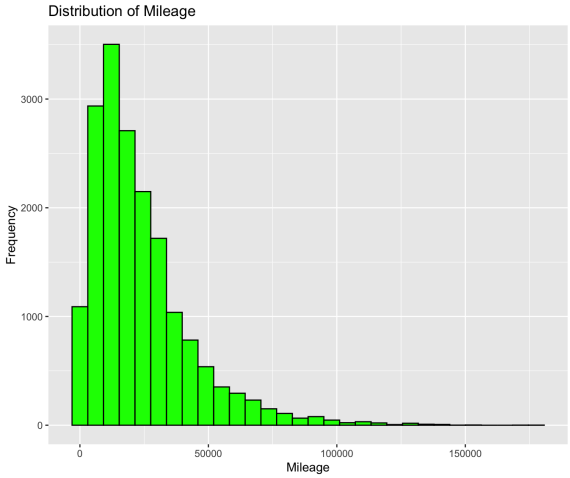
Data Cleaning

Data cleaning involved removing entries with a year greater than 2024. Additionally, a new variable, age, was calculated by subtracting the vehicle's year from the current year, set at 2024.

Exploratory Data Analysis

Before delving into complex models, an EDA was conducted to gain insights into the data. Histograms were created to understand the distribution of vehicle age, mileage, and price. The distributions provided a preliminary view of the data's shape, spread, and potential outliers.





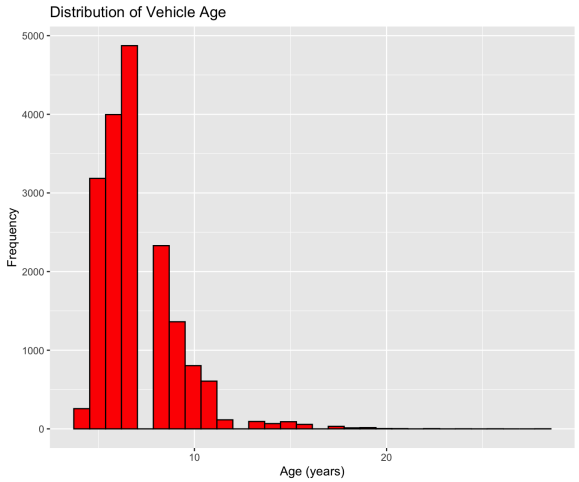


Figure 1: Histograms showing the distribution of vehicle age, mileage, and prices, revealing patterns and spread in the dataset used for regression analysis.

Scatter plots were also examined to visualize the relationships between price and the independent variables of mileage and age. These plots suggested a potential negative correlation between both mileage and age with the price of the vehicles.

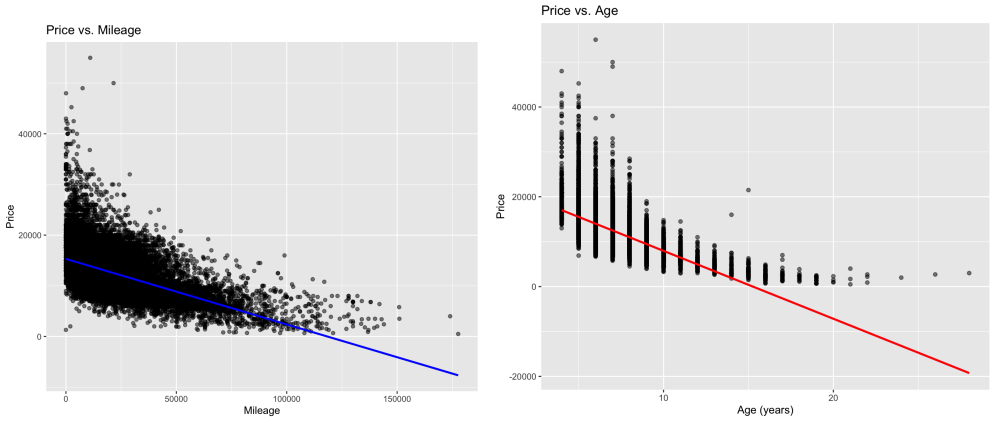


Figure 2: Scatter plots of price versus age and mileage, with fitted regression lines indicating the trends in the data.

Model Development

The WLS regression model was specified as follows:

$$Y = \beta_0 + \beta_{mileage} X_{mileage} + \beta_{age} X_{age} + \epsilon$$

Here, Y represents the price of the vehicle transformed using the Box-Cox transformation to correct for non-normality in the residuals, $X_{mileage}$ is the mileage of the vehicle, and X_{age} is the age of the vehicle. The weights in the WLS model were inversely proportional to the age of the vehicle to stabilize the variance of residuals across different age levels.

The suitability of the model was assessed using diagnostic tests, including the Shapiro-Wilk test for normality of residuals and the Breusch-Pagan test for heteroscedasticity.

The Breusch-Pagan test yielded a test statistic of 505.81 and a p-value < 2.2e-16. The test statistic is large, and the extremely small p-value suggests strong evidence against the null hypothesis of homoscedasticity. This indicates that there is significant heteroscedasticity in the model's residuals.

The Shapiro-Wilk test resulted in a W statistic of 0.98665 and a p-value < 2.2e-16. While a W statistic close to 1 suggests that the data is normally distributed, the p-value here indicates that we can reject the null hypothesis of normality.

The results indicated a need to transform the response variable, price, to meet the assumptions of the regression analysis. The Box-Cox transformation was applied to the price variable, resulting in a lambda value that maximized the log-likelihood of the model.

Hypothesis Testing:
The primary focus of the statistical analysis is to determine the impact of mileage $X_{mileage}$ and age X_{age} on the price of vehicles.

$$H_0 = \beta_{mileage} = 0$$
$$H_0 = \beta_{mileage} \neq 0$$

$$H_0 = \beta_{age} = 0$$
$$H_0 = \beta_{age} \neq 0$$

The regression output:
The coefficient for mileage is estimated at -1.685e-06 with a standard error of 8.645e-08, yielding a t-value of -19.50. The associated p-value is less than 0.001, indicating a statistically significant negative relationship between mileage and vehicle price.

The coefficient for age is estimated at -7.847e-02 with a standard error of 8.681e-04, and the t-value is -90.39. The p-value is also less than 0.001, suggesting a statistically significant negative relationship between the age of a vehicle and its price.

Given the extremely low p-values associated with both predictors, we reject the null hypotheses for mileage and age, confirming that both are statistically significant factors in determining the price of a vehicle. We can result in the following equation:

$$Price_{transformed} = 7.728 - 1.685 * 10^{-6} * mileage - 0.07847 * age$$

Model Diagnostics

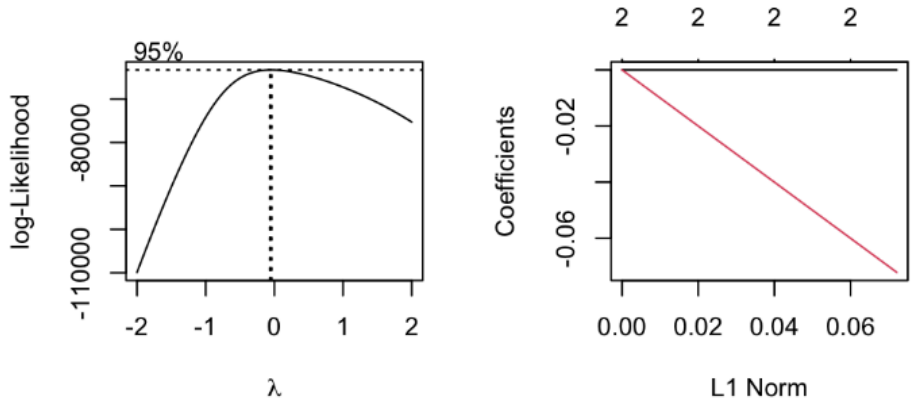


Figure 3: Box-Cox transformation plot indicating the optimal lambda value for normalizing the distribution of the response variable

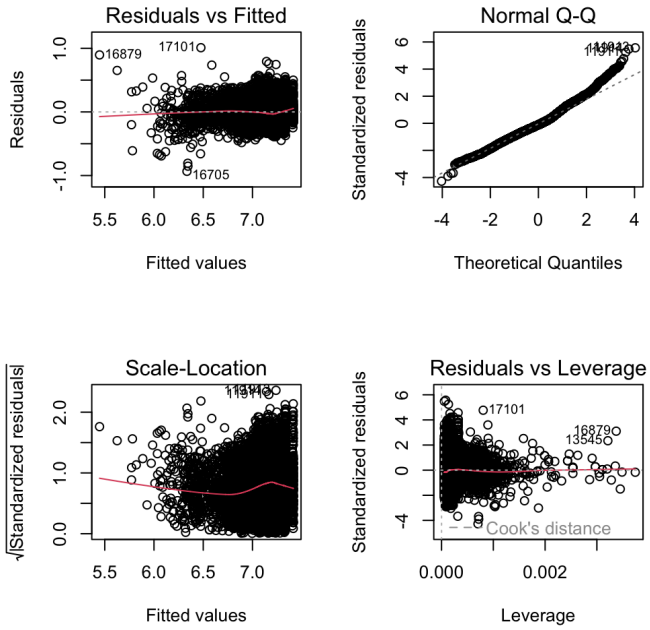


Figure 4: Diagnostic plots for the WLS regression model assessing the assumptions of homoscedasticity and normality of residuals.
A Box-Cox transformation with a lambda of 0.4545 was initially attempted but did not resolve the diagnostic issues. The subsequent application of the WLS regression model resulted in an improved fit, with an R-squared of 0.3568 and adjusted R-squared of 0.3566.

The diagnostic plots further indicated a satisfactory model fit, with no obvious patterns in the residuals, and a near-normal distribution, suggesting that the transformed model had addressed the initial concerns of heteroscedasticity and non-normality of residuals.

Conclusion:
The analysis conducted provides clear evidence that both mileage and age significantly influence the price of Ford vehicles. The negative coefficients for both predictors indicate that higher mileage and older age are associated with

lower prices. However, the model diagnostics highlight underlying issues with the data that could affect the reliability of the regression results.

10 fold cross examination:
I implemented a 10-fold cross-validation to evaluate the model's robustness. The original model, untransformed, revealed an RMSE of 3591.364 and an R-squared value around 0.4267, explaining roughly 42.67% of the variance in vehicle prices, which is an acceptable fit. However, the introduction of the Box-Cox transformation gave significant enhancements. The RMSE went down to 0.1435874, while the R-squared went up to 0.61759, suggesting the model now captures around 61.76% of the variance in the transformed vehicle prices. The MAE also improved, decreasing to 0.1109057.

These improvements in model metrics, while impressive, necessitate a nuanced interpretation. The lower RMSE and MAE values, post-transformation, reflect the altered scale of the dependent variable and not necessarily an intrinsic improvement in predictive accuracy. To genuinely comprehend the model's predictive enhancements, I would reverse the Box-Cox transformation on the predictions and reassess the RMSE and MAE against the original price values. Nonetheless, the increased R-squared value bolsters the validity of our model adjustments, affirming that our transformed model better adheres to the underlying principles of linear regression and more accurately reflects the data's intricacies.

Limitations
The dataset may be missing influential variables beyond mileage and age—such as vehicle condition, included features, and prevailing market trends—that could impact the pricing of vehicles. Additionally, despite our efforts, the model's residual diagnostic checks show some underlying issues, which could undermine the validity of our statistical tests and confidence in the intervals calculated.

Furthermore, the attempted Box-Cox transformation, while promising in theory, fell short of completely getting rid of heteroscedasticity and normality within the residuals. This shortfall underscores the potential necessity for alternative transformations or even more robust modeling strategies to capture the complexity of vehicle pricing accurately.