A Multiple Regression Analysis of Prices on Used Electric SUVs

Study Conducted on Ads from blocket.se

from Car Dealers located in the Skåne Region

Astrid Hansen

EC Utbildning

Kunskapskontroll – R Programmering

202404

Abstract

This report centers on conducting a multiple linear regression analysis of prices for used electric SUVs sold by car dealers in the Skåne Region. The objective of the study is twofold: firstly, to identify a suitable model for price predictions to assist car dealers of electric SUVs, and secondly, to explore inference to enhance customer awareness of the factors influencing prices.

The study findings indicate that horsepower exerts the biggest influence among the selected variables on price setting, whereas mileage exhibits a smaller influence than anticipated. Conversely, the color variable has the least effect, while the car brand also demonstrates some influence. The final model's predictions could be enhanced, and the model fit could be improved with a more complex model. However, for interpretability reasons, a simpler model was preferred.

Contents

1 Introduction

Data from SCB on new car registrations by fuel type demonstrates a notable pattern: the market for electric vehicles is experiencing substantial growth. Additionally, there is a decrease in the number of registrations for fossil-driven cars. The graphs below illustrate the changes taking place over a relatively short period, from 2020 to 2023 in Sweden.

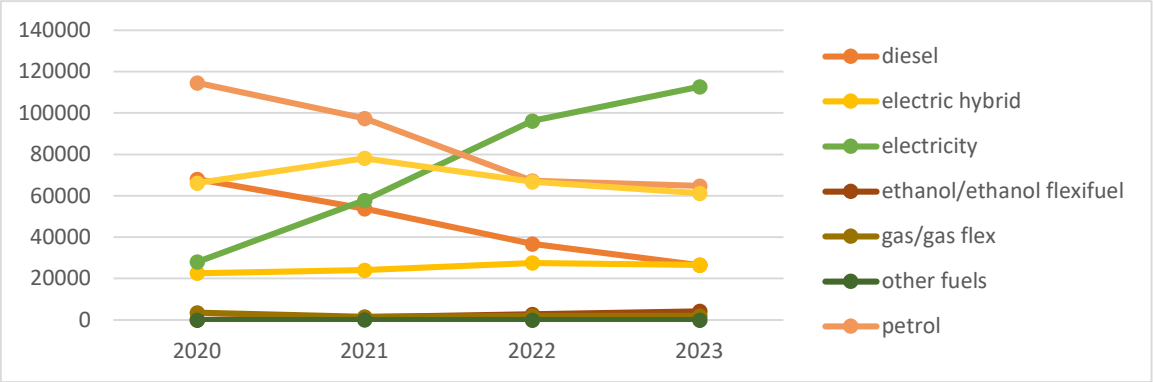

Figure 1: Line chart over new car registers in Sweden per fuel type 2020-2023

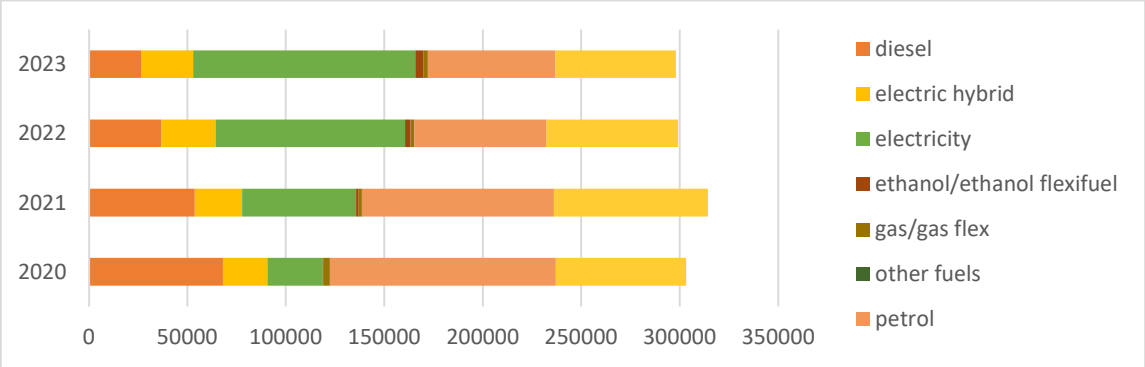

Figure 2: chart over new car registers in Sweden per fuel type 2020-2023

Inspecting the numbers behind the charts then the increase for registration of new electric cars goes up from 28.097 in 2020 to 112.775 by the end of 2023. An increase of 301%. The trend in the chart shows that the environmental agenda to reduce CO2-levels is prominent in society, evident both in public consciousness and legal frameworks, as well as in consumer behavior. The EU's legislation to ban the sale of new gas and diesel cars from 2035 emphasizes and encourages the growing presence of electric and other non-fossil fuel vehicles on our roads.

The trend emphasizes the importance of analyzing the market for these cars. Such analysis is essential to aid car dealers in setting fair selling prices as well as consumers to comprehend which factors drive price levels. This is particularly important because the factors influencing the valuation of a used electric car can differ from those affecting a gasoline-driven vehicle.

1.1 The Aim of the Study and Questions

The considerations mentioned above raise the questions that this report seeks to investigate and reflect upon. The aim of this study is to find a model that can predict the selling price for used electric SUVs set out for sale by car dealers in the Skåne region on the website blocket.se, as well as

examining the factors that influence the prices and to what extent. To investigate these questions, I will address the following:

1. What are the characteristics of our dataset, and how can we use it in a regression analysis?
2. Which regression model provides the most accurate representation of the relationship between the variables and yields the best price predictions?
3. What is the relationship between the response variable price and the predictor variables' mileage, horsepower, car brand, color, and model year?

1.2 Conceptual Framework

The variables selected for this study, which are related to the dependent variable "price," include brand, model year, mileage, horsepower, and color. These variables were chosen based on their prominence in car advertisements on Blocket, indicating their importance for customers in making purchasing decisions.

It is anticipated that all these variables will have an impact on the price, with mileage and horsepower expected to exert the greatest influence (negative and positive), while color is expected to have the least impact. Traditionally, these variables are regarded as reliable indicators of price for gasoline-powered vehicles, as they reflect both usage (and consequently wear and tear) and engine power.

1.3 Limitations of the Report and Reflections

The study is limited to examining a few selected variables. Other relevant variables, such as battery life, specific car models, and drive per charging, would have been interesting to investigate and could be explored in future studies. But lack of time and capacity put a natural limitation on the data collection.

The decision to exclusively analyze used cars listed for sale by car dealers in the Skåne region was motivated by the intention to construct a model that aids local car dealers in benchmarking their prices against competitors. This decision rests on the premise that prices for electric cars may fluctuate across regions due to variations in charging infrastructure and electricity prices. Additionally, it is assumed that cars sold by private individuals tend to be priced lower.

The limitation of focusing on SUVs stems from their prominent presence on Blocket, where there is a substantial number of electric SUVs compared to estate and hatchback cars. These limitations were imposed to obtain a dataset that is representative of a specific area and to avoid working with a dataset that is imbalanced. Whether these limitations are too narrow will be touched upon during the study.

1.4 Report Structure

The first section covers the theory behind the study, including discussions on data collection, variable handling, as well as models and model measures. The next section focuses on the methodology and describes the procedures implemented in the code. The last section focuses on results and discussions of the findings, before concluding with a summary and addressing the initial questions posed.

2 Theory

This section outlines the theoretical framework behind the methodology employed to investigate the study objectives.

2.1 Multiple Linear Regression

The primary focus of this study is to conduct a multiple linear regression analysis on the collected data. The rationale for this approach stems from the expectation that the relationship between the response variable and the predictor variables can be adequately represented by a linear model. One of the key advantages of a linear model lies in its simplicity and interpretability. It's worth noting that while non-linear models may offer better predictive performance, they are often more complex and challenging to interpret statistically (James, p. 2020).

Multiple linear regression is a parametric method, as it assumes a linear relationship between variables, unlike non-parametric methods. Consequently, in multiple linear regression, our task involves estimating the coefficients of the regression line rather than the entire function (James, p. 21). Given that this study aims to both develop a predictive model and conduct statistical inference analysis, opting for a linear model seems appropriate.

2.1.1 Assumptions of the Linear Regression Model

Several assumptions underpin the linear model. These include homoscedasticity, which posits that the error variance remains constant across all values of the independent variables. Furthermore, the model assumes independence of observations, meaning there are no concealed relationships between them. Lastly, the data is expected to have residuals that follow a normal distribution.

Should the data fail to meet these assumptions, further considerations may be necessary. This could involve acquiring additional data to address deficiencies or opting for a nonparametric test/model as a more suitable alternative.

2.2 Model Evaluation and Measures

For both simple and multiple linear regression, the goal is to find coefficients in the model that minimize the sum of squared residuals (RSS). The coefficients for a variable in a simple regression can deviate quite substantially from the same variable placed in a multiple linear regression model together with other variables (James, p. 73).

2.2.1 F-Statistic and p-values

If a models F-statistic is close to 1, then it indicates that there is no relationship between the response and the predictor variables. The interpretation of the F-statistic depends on the ratio of variables to observations. The larger the dataset relative to the number of variables, the smaller the F-statistic needs to be for us to reject the null hypothesis (James, p. 78). The null hypothesis states that none of the predictors in the model are related to the response variable, while the alternative hypothesis suggests that at least one of them is not equal to zero (James, p. 78). Along with the F-statistic, a p-value is provided, indicating whether we can reject the null hypothesis as described

above. The p-value in this context is calculated with the threshold set at $p < 0.05$ (95% confidence interval) to reject the null hypothesis.

Given our limited number of observations and variables, the F-statistic serves as a suitable measure for testing the association between the predictors and the response variable (James, p. 78).

2.2.2 Measurements

One model evaluation metric is the RSS, which stands for the residual sum of squares. It measures the variance of the residuals to the regression line and should preferably be as low as possible. Another measure is the RSE which measures the residuals standard error and are an indication of the variance of the residuals to the regression line.

Another measure is R^2 that is a measure of how much of the variance in the response variable is explained by the model. This should in turn be as high as possible. Adding more predictors increases the model's explanatory power on the training set, as well as the R^2 but it doesn't necessarily mean that the model performs well on unseen data. To mitigate the risk of overestimating the model's performance indicated by R^2, the adjusted R^2 is introduced. Adjusted R^2 penalizes an increase in predictor variables in the model and should therefore be used when evaluating models with multiple predictors. We aim for a high adjusted R^2 since it represents how much variance is explained by our model (James, p. 235).

Other evaluation metrics include Cp, AIC, and BIC, which adjust the training error to provide an estimate for the test error, aiding in model evaluation. There are slight differences in their measurement approaches, with BIC favoring simpler models compared to AIC and Cp. However, small values in all these measures suggest a potentially low test-error in the model.

Another measure for evaluating a model is RMSE (root mean squared error), that gives insights into the model's predictive abilities. It is not to be mistaken for the RSE that are a measure of variance of the residuals. But both measures have in common that lower values implies a better model.

2.3 Problems with Linear Regression Models

When opting for regression with a linear regression model, we rely on assumptions regarding the relationships between the response and predictor variables, as discussed previously. Failure to meet these assumptions can lead to potential issues in estimating this relationship. The following section delves into some of these challenges.

2.3.1 Non-Linearity

If the assumption of linearity is violated in a linear regression model, it poses a significant problem. In cases where the relationship is non-linear, alternative regression models or variable transformations should be considered to accurately represent the relationship. Non-linearity can be identified through residual plots. When multiple predictors are involved, plotting the residuals against the predicted values can reveal patterns that signal non-linearity in the relationship between the response variable and the predictor variables (James, p. 93).

2.3.2 Correlation of Error Terms

Correlation between error terms can distort the calculation of standard errors, resulting in underestimated values for prediction and confidence interval, as well as for p-values, potentially leading to inflated model assessments (James, p. 94). The correlation of error terms arises from a violation of the assumption that observations in the dataset are independent of each other. This issue can be mitigated through careful experimental design before data collection. Detecting correlation between error terms can be difficult. It is typically a problem in time series data or when observations share similar environmental exposures, leading to correlated residuals.

2.3.3 Non-Constant Variance of Error Terms

Non-constant variance of error terms, also known as heteroscedasticity, can impact interval calculations and the model's significance. Ideally, error terms should exhibit homoscedasticity, characterized by constant variance. To assess problems with non-constant variance - heteroscedasticity - one should examine a plot of residuals against predicted values and look for a funnel shape. The funnel shape indicates a widening in variance. Transforming the response variable may help address this issue (James, p. 96).

2.3.4 Outliers

Outliers are observations that have a substantial deviation from the predictions of the model and are commonly encountered in datasets. The problem arises when outliers affect the model's calculations – has high leverage. If an outlier exhibits "normal" values for the predictor variables, it typically has minimal influence on the least squares fit of the model. However, it can still impact the Residual Standard Error (RSE), leading to complications in interval and significance calculations. To diagnose whether a value deviates significantly enough to be considered an outlier, one can examine the studentized residuals as well as high leverage plots.

To address issues with outliers, the first step is to revisit the values in the dataset. This involves checking for potential measurement errors or missing predictor values.

2.3.5 High Leverage Points

These points exhibit unusual values for the predictor variables. While outliers have minimal influence on the model's regression line if they have low leverage, high leverage points can significantly affect it. Identifying high leverage points can be more challenging with multiple predictors compared to a simple linear regression model. *"...it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors"* (James, p. 99).

Detecting high leverage points is accomplished using leverage statistics. A high leverage point is suspected if its leverage statistic exceeds (**p + 1) / n, where p represents** variables and n represents observations. The threshold is the average leverage for all observations. These points can be visualized in a plot of studentized residuals against the leverage statistic. If the plots reveal high leverage points, then one could examine their influence on the model's regression line before deciding on further actions, like removal.

2.3.6 Collinearity

Lastly, the dataset should be examined for linear correlations among the predictor variables. If predictors are correlated, it becomes challenging to draw inference analyses due to uncertainty regarding the influence of correlated predictor variables on the response variables. The presence of collinearity leads to uncertain coefficient estimates and inflates the standard errors of the coefficients, thereby impacting the t-statistic and hypothesis testing outcomes (James, p. 101).

Collinearity in the dataset can be identified through line plots or correlation matrices. However, it's important to note that multicollinearity can exist between three or more variables even if no pair of variables has a particularly high correlation (James, p. 102). The variance inflation factor (VIF) measure is a reliable method for assessing collinearity. VIF compares the variance in the coefficient estimate in the full model to the variance of a coefficient estimate in a model with only that predictor. A VIF value of 1 indicates no collinearity, while values over 5 could indicate problematic collinearity.

The VIF measure can be applied to models with both categorical and numerical values, as found in our dataset. For categorical predictors, collinearity can be evaluated using the generalized VIF (GVIF) score, calculated as GVIF^(1/(2*Df)). GVIF^(1/(2*Df)) should be squared whereafter it can be evaluated according to the standard VIF rule of thumb.

To address collinearity, one can either remove one of the correlated predictors or combine them into a single variable. However, collinearity between predictors is not always problematic and may be disregarded in certain scenarios. For instance, categorical predictors with a small number of observations for some values may result in higher variance and VIF values without significantly impacting the model fit. Additionally, if one of the correlated variables with a high VIF is considered a control variable and the other correlated predictor of interest has a low VIF, collinearity may be deemed acceptable (Allison, P. 2012).

2.4 Statistical Inference

Our objective for this study is twofold: to develop a model that achieves accurate predictions while remaining simple. Simplicity facilitates the assessment of the influence of predictor variables on the response variable, which aligns with our interest in statistical inference. A linear model with few predictors enhances accessibility to statistical inference and is therefore suited for this studies aim. The coefficients of the predictor variables provide valuable insights into their influence on the response variable. To identify the coefficients for the optimal regression line, we must also evaluate the model's predictive performance. However, it's important to note that measures of error rates for model predictions should not be the sole determinant in selecting the final model. Other factors and considerations should also be considered.

3 Method

This section covers the methods used to investigate the aim of the study. It will include the data collection process, the investigation and transformation of variables, the model training conducted, as well as the evaluation of different models and assessment of potential problems.

3.1 Data Collection

As described in the introduction, the focus of the study is quite narrow, centering on used electric SUVs listed for sale in the Skåne region by car dealers. This approach serves two purposes: addressing a potential business case and minimizing the risk of obtaining a dataset that is too unevenly distributed, which could complicate the regression analysis. It is challenging to obtain sufficient data that can serve as a representative sample without including too many variables with insufficient representation of certain values.

The goal was to obtain between 450 and 500 observations within the study's scope. Ads were selected up until the 10th of April and can be regarded as a sample of the population of used electric cars listed for sale by car dealers in the Skåne region. Since ads change continuously due to cars being sold and added, randomization in the data collection was achieved by selecting the time frame (ads up until the 10th of April). The variables that were collected are described in the following section about the data set. Seasonal variations in prices are beyond the scope of the study.

In collecting the data, we opted to focus on SUVs due to their prevalence in advertisements for electric cars. Additionally, we were interested in the listed price of used cars rather than new ones, which led us to filter out cars that had driven less than 50 miles.

3.1.1 Limitations

Almost all electric SUVs have automatic gearbox, so that variable is not interesting to investigate. Drivetrain, that is also listed on blocket as one of the features, are also excluded. The hypothesis is that it does not have a big influence on the price in relation to the other variables. As the aim is a precise model, but also a simple model, then this variable has been left out. In hindsight, this could though have been an interesting variable as well.

Other likely important features include battery life and drive per change which would have taken longer time to gather information about. The car model may also influence the price but is excluded from this study to prevent too few observations for certain models. Other features such as car conditions and accessories do also play a role in setting the right price, but as they are most likely never the ones that holds the biggest influence for the price, and because of the extra manual work in extracting this information as well, then they are not part of the study.

3.1.2 SUV

A main question to address is also: what constitutes an SUV? SUV stands for Sport Utility Vehicle and encompasses a wide range of cars. They are not classified by Transportstyrelsen as a distinct car type, so the definition of an SUV can vary significantly from one market to another. Generally, an SUV is considered a robust vehicle with extra space and a higher seating position. Its popularity has grown over the years, as evidenced by the numerous SUVs sold on Blocket.

3.2 Data Import and Data Cleaning

Before importing the data into R, it undergoes a thorough check and cleaning process in Excel. This involves removing both visible and hidden blank spaces, ensuring that numeric values are formatted correctly, verifying that all numeric values can be sorted in the correct continuous order,

and checking and correcting text variables to eliminate duplicates (e.g., "Blå" and "blå") and detect any misspellings.

After cleaning and checking the data in Excel, it is imported into R as an excel file. The imported data set can be inspected with several R functions, which is done to verify that it is imported correctly. The dataset consists of 467 observations and 6 variables. The brand and color columns are labeled as character data types, while the numeric columns (mileage, horsepower, model year and price) are labeled as 'dbl'. Since this study is limited to this dataset and does not involve any decimal values, I will convert the numeric columns to integers. During this conversion, the numeric data is checked for blank spaces again. Furthermore, the character variables will be transformed into factors later in the code, enabling their use in our models.

3.3 Exploratory Data Analysis

The goal of the exploratory analysis is to investigate the distributions and correlations of the variables, as well as to lay the groundwork for possible transformations of the variables.

3.3.1 Numeric variables

As stated above, there are 4 numeric variables in the dataset. They are listed below along with their range, mean, and 1st and 3rd quartile values.

Variabels	Min	Max	Mean	1st Quantile	3rd Quantile
Price	199.000	1.599.900	518.588	429.850	589.000
model_year	2016	2024	2022	2021	2023
Mileage	50	18.900	3074	950	4246
housepower	136	1020	287	204	346

Figure 3, table with numeric variables

3.3.2 Response Variable - Price

In the table above, we observe that 50% of the values in the 'price' variable fall within a relatively narrow range, from around 430,000 to 589,000. Examining the distribution of the variable, we notice that most values are centered around 500,000, with a right tail indicating a few observations with high price levels. These could potentially be outliers in the dataset, which may affect the model's representation of the relationship between the variables.

Ideally, the response variable should be normally distributed or symmetric which would simplify the process of determining an effective explanatory model. With that said, then it is not a demand that the response variable is normally distributed to do a linear regression model.

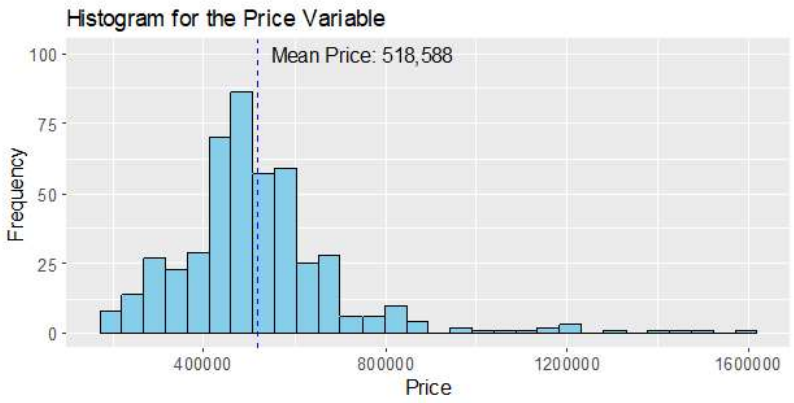

Figure 4: Histogram for the price varaible

3.3.3 Model Year

The predictor variable 'model Year' ranges from 2016 to 2024, with most data points labeled from 2021 onwards. This is not surprising, given the negligible number of electric cars on the market in earlier years. However, the small number of values in the early years can pose problems for our model, so we group them accordingly. Considering the rapid evolution and improvement in the manufacturing of electric cars, it is reasonable to assume that prices vary from year to year Therefore, the variable is divided into cars before and including 2020, 2021, 2022, 2023, and 2024, instead of three broader categories: old, middle, and new.

Running a simple linear regression on model year (the ungrouped variable) and price also indicates an increase in prices for each year, starting from around 2021 and upwards. The new year_category variable is lastly changed to a numeric data type to be able to keep the ordinal nature of the values.

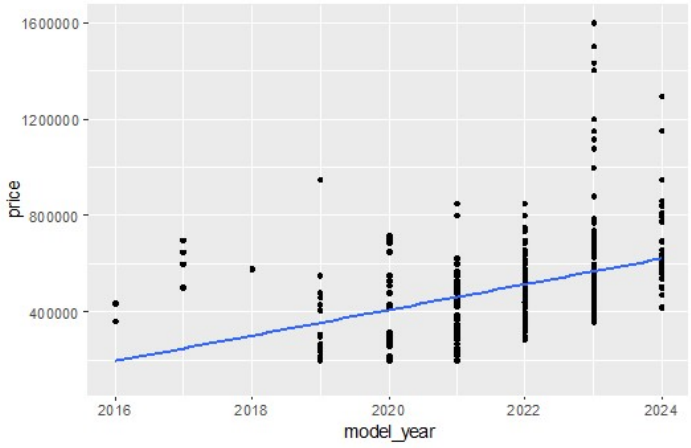

Figure 5: model year variable and price

One could point out the problem of grouping the year category with one value covering “old” cars and the other as normal years. But the few observations up until 2021 will pose problems for our model. Even with this grouping, the “old” car category ends up being the smallest one in the variable with 54 observations. Another approach could have been to remove them all together and focus on cars after 2021.

3.3.4 Horsepower

Horsepower will be treated as a continuous variable, although it could also be grouped into sections. It ranges from 136 to 1020, with a mean of 287. A histogram indicates that many cars have horsepower around 200, while very few have horsepower over 450. There also appear to be potential high leverage points, with a value of 1020 in the dataset, which will be investigated further in the study.

A plot reveals a positive correlation between horsepower and the price variable, which is not surprising. However, there is also a significant amount of variance around the linear regression line.

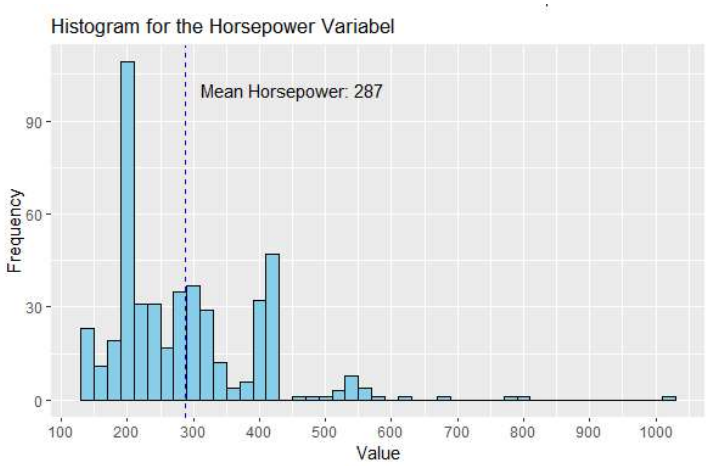

Figure 6: Histogram of the Horsepower variable

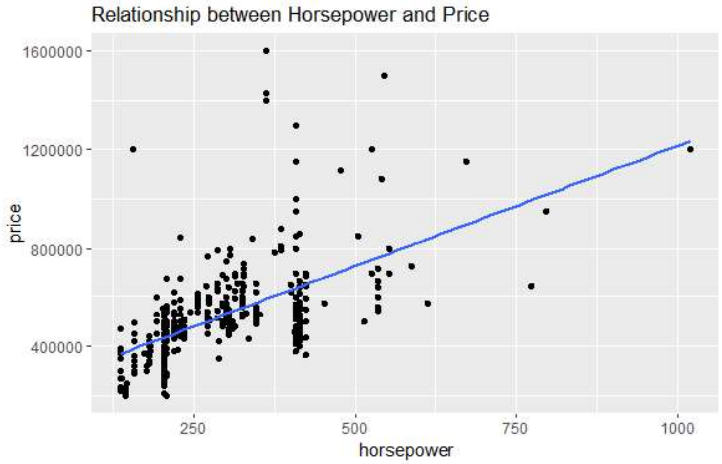

Figure 7: Plot between horsepower and price

3.3.5 Mileage

A histogram of the mileage variable reveals a left-skewed distribution, with many values clustered around 1000-1500 km and a mean of 3074 km. Additionally, there appears to be a potential high leverage point, with one data point having a value of 18,900 km, which greatly deviates from the other data points.

When plotting the relationship between mileage and price, we do not observe a clear negative linear relationship as expected. While there is a negative correlation, it does not seem to

follow a linear pattern. As a result, it could be tested to either transform the price variable or the mileage variable.

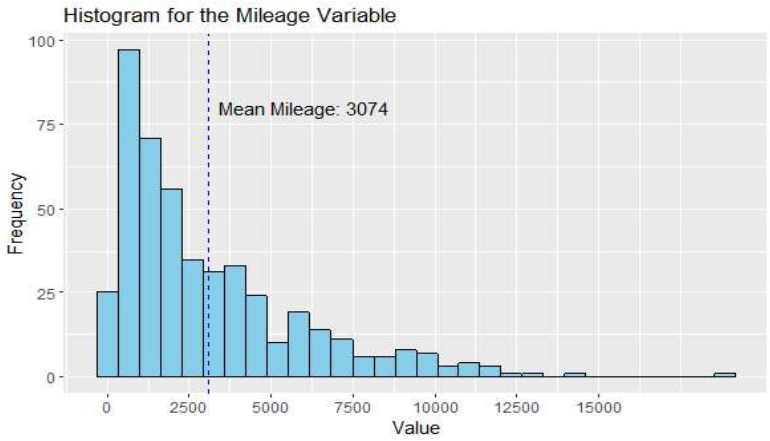

Figure 8: Histogram of the Mileage Variable

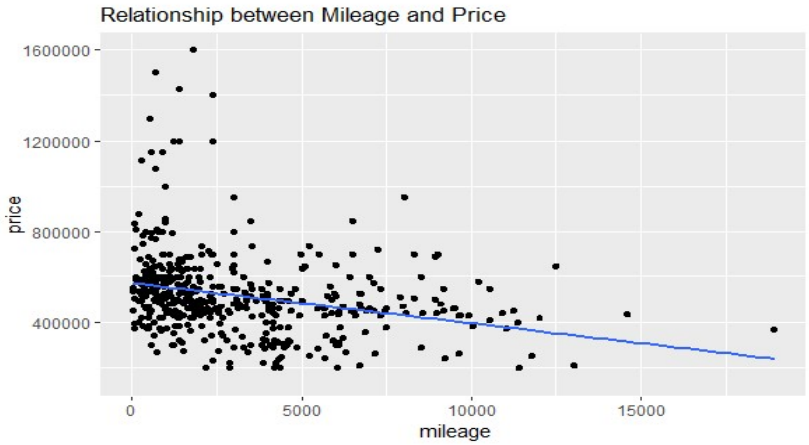

Figure 9: Plot between mileage and price

3.3.6 Brand

Given the widespread attention, particularly surrounding Tesla cars, one would anticipate some form of brand or car model influence on pricing. However, analyzing this variable proves challenging due to the limited number of observations for certain car models and brands.

To gain preliminary insights into potential brand effects - understood here as a more subjective measure encompassing consumer emotions and perceptions-, I have opted to retain brands represented by over 20 data points in the dataset, grouping all others under the category "other." It's worth noting that this approach may yield some degree of ambiguity, as the "other" category encompasses a diverse range of brands, including both popular and luxury options such as Jeep and Jaguar, as well as more commonplace brands.

Reviewing the boxplot with the updated brand categories it becomes apparent that the "other" category encompasses brands displaying some price deviations, but also that it is centered around the mean of the price variable. The plot also shows variations in the prices of the car brands,

with Kia positioned at a lower price point compared to brands like Tesla, Mercedes-Benz, Audi, and BMW, which could be considered top-tier options.

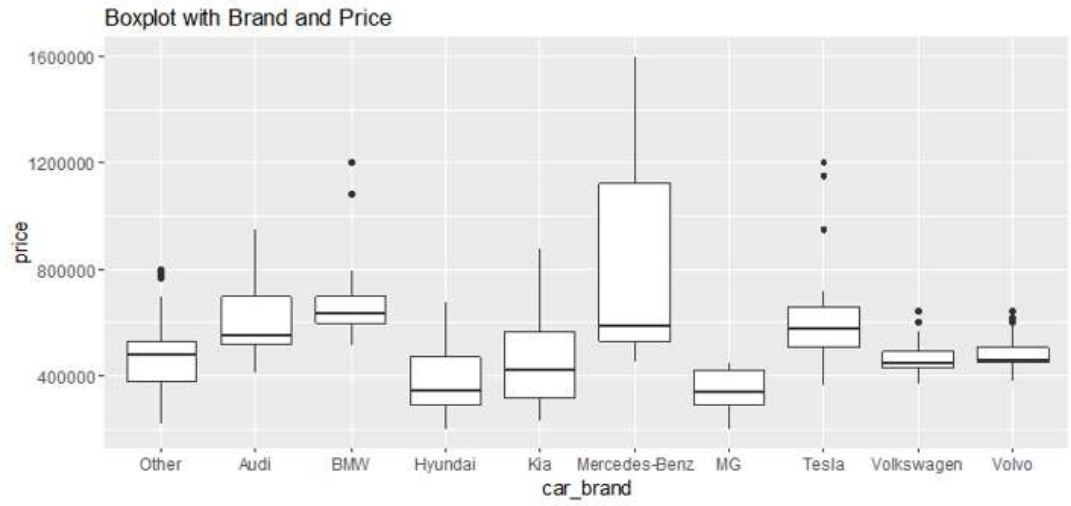

Figure 10: Boxplot with the brand category and price

Grouping variables together can be a useful approach, but it is essential to reflect on the groupings and preferably conduct a sensitivity analysis and test the model both with and without the groupings. The advantage of grouping uneven distributed variables is that estimates based on few data points tend to be imprecise.

3.3.7 Color

The last variable is the color variable. In the introduction, this variable was assumed to have the least influence on the price settings. When checking the variable, it is evident that some colors are represented more frequently than others, necessitating the grouping of the variable due to insufficient data points for certain colors.

The revised variable is now categorized into Blue, Gray, Other, Red, Black, and White. A box plot illustrating the relationship between these new color categories and price indicates that most cars categorized by color fall within a similar price range. Nonetheless, there are data points for some colors that exhibit higher price values.

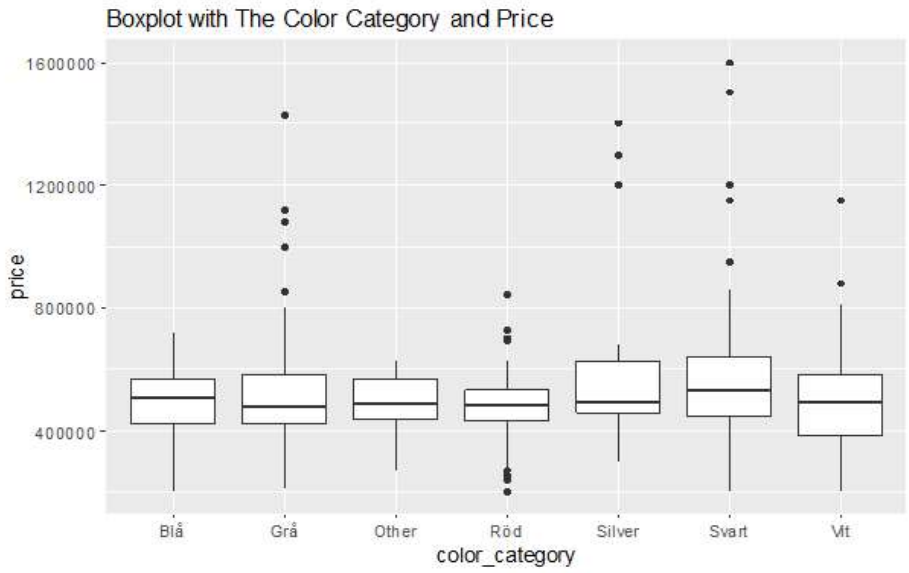

Figure 11: Boxplot with color_category vs price

3.3.8 The Updated Data Set

The transformed dataset consists of 6 variables. The factorial (qualitative variables) will be handled as dummy variables in R when we run our regression model (James p. 119).

mileage	int
horsepower	int
price	int
year_category	int
car_brand	Factor / 10 levels
color_category	Factor / 7 levels

Figure 12: Table with variables

3.4 Correlations

Before proceeding with model training and testing, the data underwent a check for correlations between the variables.

In the correlation matrix in figure 13 we see a negative correlation between the mileage variable and the year_category, which could pose a problem for our model. The correlation between the horsepower and price variable is the one that shows the clearest sign of correlation in the plots, while the correlation between the mileage and horsepower variable is small.

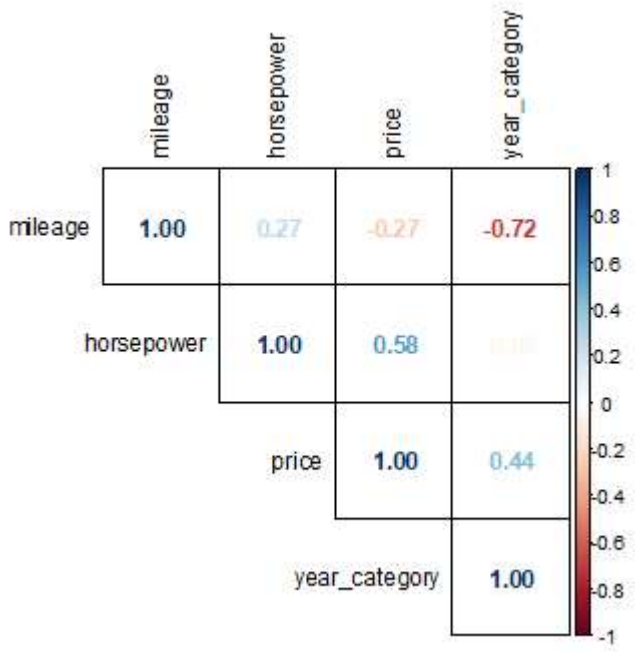

Figure 13: Correlation Plots

3.5 Model Training

To construct a model that not only offers accurate predictions but also effectively captures the data without overfitting, the dataset will be partitioned into three subsets: train, test, and validation sets. Subsequently, the chosen models will be trained using the training set, assessed using the validation set, and ultimately tested on the test set.

3.5.1 Model 1

model_1 consists of the following variables: horsepower, mileage, year_category, color_category, and car_brand. The p-value was significant, and the adjusted R^2 was 0.7236. The numeric variables were significant at a three-star level, while certain variables in the car_brand showed significance. None of the color_category variables were significant.

Assessing the VIF-values of the numeric variables reveals values below 5, indicating low collinearity. For the categorical predictors, the VIF-values are computed by squaring the GVIF^(1/(2*Df)) measure. These values suggest no substantial collinearity issues among the variables.

model_1	GVIF	Df	GVIF^(1/(2*Df))	GVIF^(1/(2*Df))^2
mileage	2,684556	1	1,638461	
horsepower	1,857119	1	1,362761	
year_category	2,687548	1	1,639374	
car_brand	3,433748	9	1,070939	1,146910
color_category	1,473489	6	1,032830	1,066738

Figure 14: table over VIF-values for model_1

3.5.2 Diagnostic Plots

Inspecting the diagnostic plots for model_1 will help identify potential problems in our dataset and with our assumptions about the linear relationship.

3.5.3 Non-linearity and Homoscedasticity

The plot below illustrates the residuals against the predicted values of the model. While there isn't a clear pattern such as a U-curve in the first plot, there is a notable amount of variance in the residuals for higher fitted values. However, given the limited number of values in that area, it does not substantially impact the line. This suggests that the assumption about linearity in the relationship can be upheld.

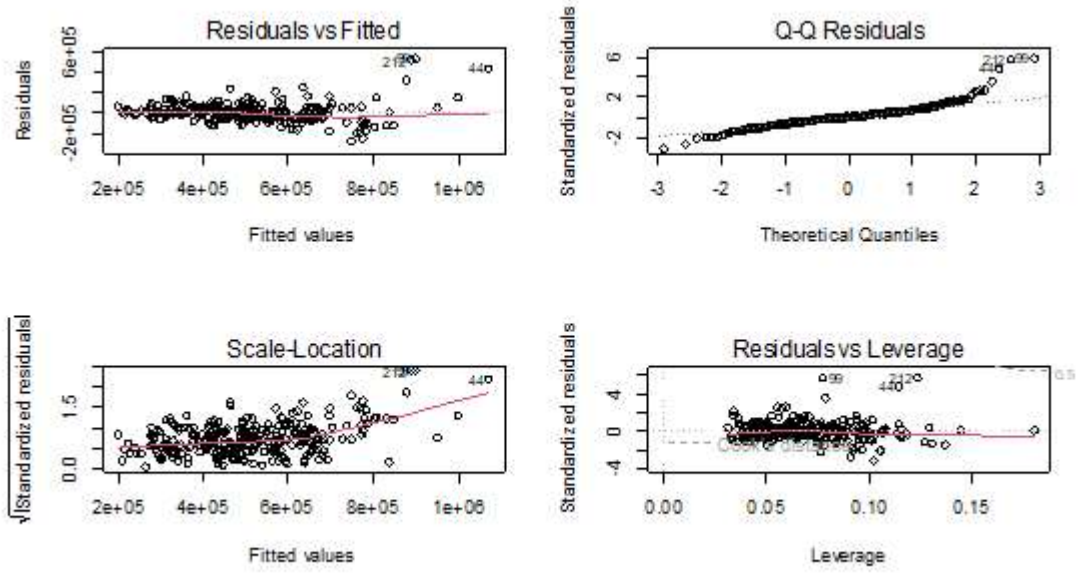

Figure 15: Diagnostic plots for model 1.

With The Scale Location plot it can be assessed whether the assumption of constant variance of the residuals can be confirmed – homoscedasticity. In our plot, the line is not perfectly horizontal, suggesting potential issues with non-constant variance. Although we do not observe a perfect funnel shape, there is a noticeable tendency for increased variance on the right side of the plot, indicating the possibility of heteroscedasticity in our model.

3.5.4 Normal Distributed Residuals

The next plot to be investigated is a QQ-plot with standardized residuals against theoretical quantiles. As explained in the theory section, we ideally expect to see a straight line, which would confirm the assumption that the residuals are normally distributed. While most of the residuals follow the straight line, there are some deviations observed for the higher theoretical quantiles, indicating potential issues with the assumption of normal distribution of the residuals.

3.5.5 Outliers and High Leverage Points

The fourth plot above shows the standardized residuals against the Leverage values. In this plot one should look for extreme values in the upper right corner or the lower right corner. The plot implies that there aren't any problematic values beyond the Cook's Distance line. The line is barely shown in the plot.

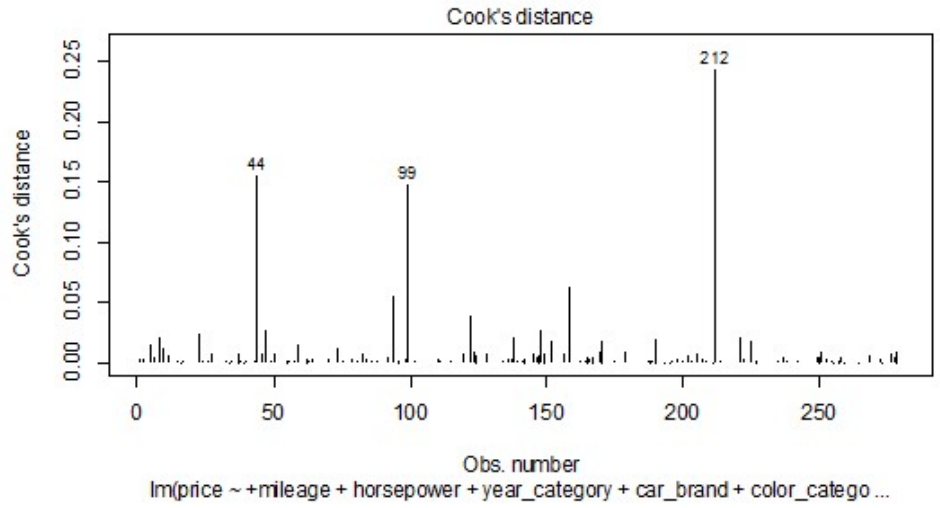

Figure 16: Diagnostic plot for Cook's Distance model_1

The plot displaying Cook's Distance reveals though a few data points at the higher end, which by inspection is observed as three Mercedez-Benz with prices over 1.4 mio. Some analysts opt for a threshold of 1 for Cook's Distance, while others utilize the formula $4/(N-k-1)$ or 4/N. Calculating the threshold using the 4/N formula for the training set yields a value of $4/280 \approx 0.014$. Employing this threshold, we can identify 18 potential problematic points in our data. Assessing the 18 data points, we see no clear pattern in the values, except for the three data points with the highest Cook's Distance.

3.5.6 Model 2 - Transformation of the Response Variable

Removing outliers and high leverage points can pose problems as they are part of the dataset and should be considered. Another solution to address the issues we diagnosed above could involve transforming the response variable. This transformation is tested in model 2 by converting the price variable to log(price). The diagnostic plot for model 2 is shown below.

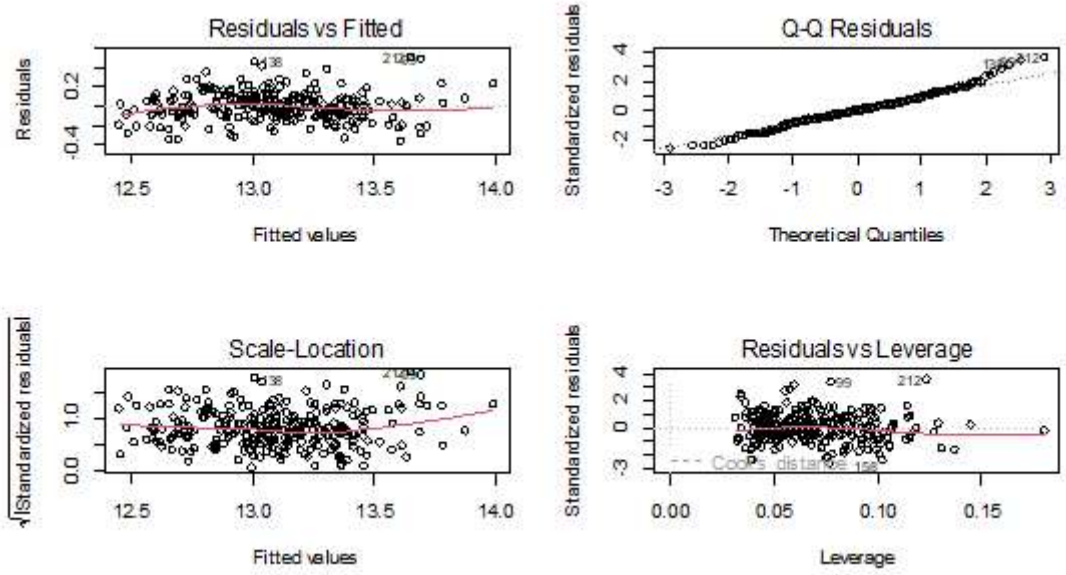

Figure 17: Diagnostic plots for model 2 with transformed response variable

There has been an improvement in addressing the non-constant variance of the residuals, and the fit between the residuals and the straight line has also somewhat improved. Overall, the improvements have enhanced the models, but they have not completely resolved all our problems.

Considering our preference for a model that is simple to interpret in terms of inference, it is preferable not to have a transformed response variable. A log transformation of the response variable alters our assumption of an additive relationship between the predictor and the expected values, instead establishing a multiplicative relationship. This change becomes apparent when, instead of taking the logarithm of Y, we exponentiate the side with the predictor variables (Cialdella, L, 2020).

Due to the modest improvement observed in the plots above and the implications associated with log transformations of the response variable, I have decided not to pursue this model further in the study.

3.5.7 Model 3 - Removing Problematic Data Points

Instead of transforming the response variable, then the 18 problematic data points found in model_1 will be removed from the training set. Model_3 will be trained with the same variables as model_1 but on the trimmed data set.

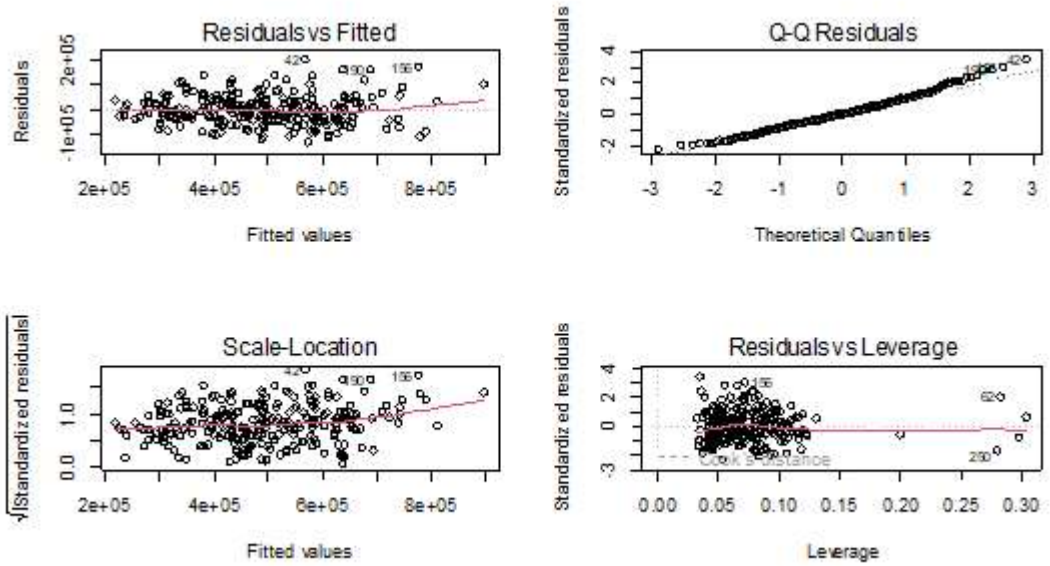

Figure 18 Diagnostic Plot for model_3

Inspecting the plots then we still see some issues with non-constant variance of the residuals, and there are new potential high leverage points. We will not attempt to remove more potential high leverage points from the data set, since we already have removed 18 from the original training set. The residuals in the QQ-plot have improved from model_1 and are now closer to the straight line.

3.5.8 Best Subset Selection

To improve the model, and hopefully get even better diagnostic plots, then a best subset selection is done on model_3 n the trimmed dataset. Runing the selection algorithm we see that horsepower is the variable to be included if we ran a simple linear regression. Looking at the model with the highest adjusted R^2, we find that it consists of 12 variables. This high number is because all our factorial variables are turned into dummy variables. We therefore have to inspect which ones are included in the selection. It turns out that there is a representative for each qualitative variable in the best 17 variable mode.

Evaluating the best subset model with the BIC value we get a model on 10 variables, where only one of the color categories is included. We did see that the color variables weren't significant when running model_1, so it could maybe improve the model by removing them, even though one of the dummy variables in the best subset model suggested by the BIC-value, were included.

3.5.9 Model 4 – Without the Color Category

The last model to be tested is the one without the color category. The model yields an adjusted R^2 of 0.8116, quite like the model with the color category. There was not a substantial change in the diagnostic plots when removing the color_category variable.

There is still a problem with a curved line in the Scale-Location plot indicating slight heteroscedasticity, and there are also some potential high leverage points as well. The distribution of the residuals is improved from the initial model_1.

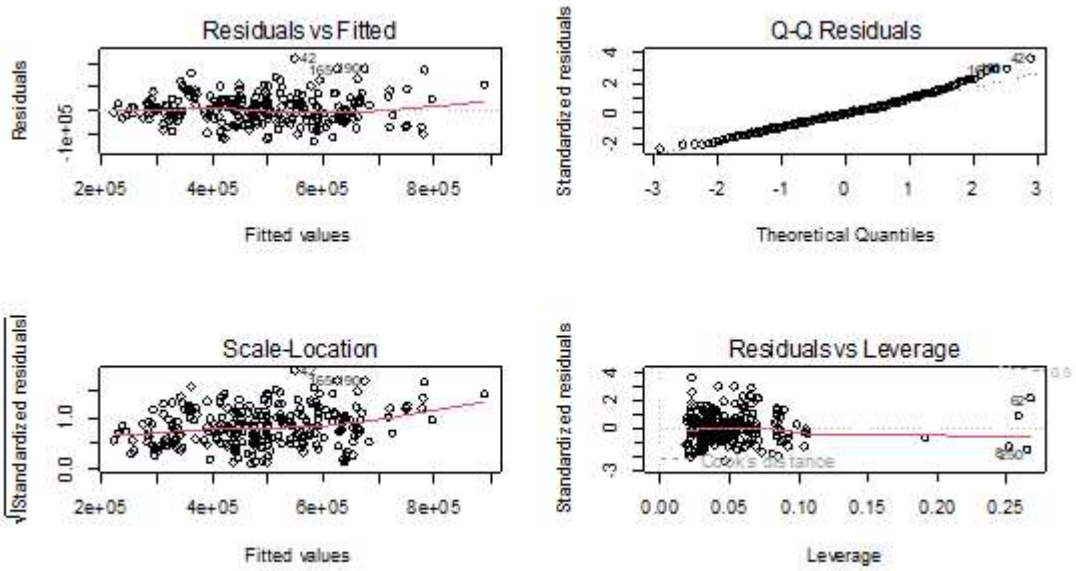

Figure 19, diagnostic plots for model_4 without the color category

3.6 Predictions

After training the model, then their predictive power is evaluated on the validation set. Below is a summary of the models that were trained and selected performance values. Since model_2 with the transformed response variable is left out of the study, then predictions will not be made for this model. All the models are significant, so we can reject the null hypothesis. There is a notable difference between model_1 and the other models, both in the adjusted R^2 and in the residual standard error, indicating that it had positive effect to remove these data points.

Models	RSE	adj.r2	p-value
model_1	94630	0.7236	2.2e-16
model_2	0.1484	0.7931	2.2e-16
model_3	60210	0.8168	2.2e-16
model_4	61060	0.8116	2.2e-16

Figure 20 Table over model values from the training set

Running the models on the validation set reveals quite similar outcomes. The Bayesian Information Criterion (BIC) is lower for model_4, which also excludes the color_category variable. Additionally, the adjusted R^2 is slightly higher in model_3, indicating that it explains variance slightly better than model_4. However, the root mean square error (RMSE) for model_4 is slightly lower than that of model_3, suggesting a better model fit. Overall, the two models are very similar in terms of prediction performance.

Model	RMSE	Adj_R_squared	BIC
model_1	107708.5	0.7235900	7303,927
model_3	110788.9	0.8168306	6602,121
model_4	109030.7	0.8116083	6582,467

Figure 21: table over predictions on the validation set for the trained models.

Since the goal was to find a simple model, then the above measure points to model_4 being the best choice. Before deciding on which model to use for investigating inference, model_3 and model_4 is tested on the test set. We get an RMSE on 110771.9 for model_3 and slightly lower RMSE on 110128.5 for model_4. Because of the lower testing error in the predictions and its simplicity then model_4 will be used for the inference study which is discussed in the following section.

4 Results and Discussion

In this section, we examine the chosen model, conducting predictions and establishing confidence intervals. Additionally, the statistical inference of the model is investigated.

4.1 The Final Model

Model_4 had 4 predictor variables: mileage, horsepower, year_category and car_brand. The RMSE on the test data can be considered ok for the purpose of this study. Considering that most prices fall into a quire narrow area surrounding 500.000 SEK, then a smaller RMSE would have been preferable if our goal was solely predictions.

Inspecting the models' coefficients we get the following regression line for a car with the car_brand Audi: Y = - 104.800.000 – 8,632 (mileage) + 816,7(horsepower) + 51.960 (year_category) + 86.950 (car_brandAudi)

Inspecting the significance levels of the variables of interest in the model, we observe a three-star significance for the mileage, horsepower, and year_category. Additionally, several car brands show significance, although not all of them. From this, we can infer that these variables indeed play a role in determining the prices of the cars on Blocket in our model.

4.1.1 Interpretation

The coefficients represent the average change in prices with a one-unit increase in a predictor variable, holding all other variables constant. For instance, the price decreases by -8,632 SEK for a one-unit increase in the mileage variable, while it increases by 816,7 SEK for a one-unit increase in the horsepower variable. Although these changes may seem small, it's important to note that mileage and horsepower can vary significantly from car to car and has relatively large ranges, especially the mileage variable.

For cars belonging to the Audi brand, the coefficient is 86.950. This coefficient is interpreted as the difference in the average prices for the car_brand Audi compared to car_brand Other. It indicates that Audi cars have an average price increase of around 87.000 SEK compared to cars categorized under "Other", which serves as our baseline category. The year variable has a coefficient of 51.960 which corresponds to the price increase the newer the car is.

The estimate for the intercept seems considerable large on the negative scale. It is not an impossible value for the regression line, but somewhat surprising. Doing the prediction and confidence intervals for different cars will give more insights into the models descriptive powers.

4.1.2 Confidence and Prediction Intervals

To investigate the prediction and confidence intervals posed by the model then we will predict the prices for these two cars.

Car_1: Mileage = 5000, horsepower = 150, year_category = 2022, car_brand = Kia

Car_2: Mileage = 1000, horsepower = 450, year_category = 2023, car_brand = Audi

	confidence_intervals			
	fit	**lwr**	**upr**	**diff**
car_1	331.165	308.841	357.490	48.649
car_2	782.071	753.481	810.660	57.179
	prediction_intervals			
car_1	333.165	210.460	455.870	245.410
car_2	782.071	658.450	905.692	247.242

Figure 22, confidence, and prediction intervals for the two cars

The average price for car_1 is 331.165, and the average value for this kind of vehicle falls in 95% of the cases in the range between 308.841 and 357.490. For this car the predicted interval is between 210.460 and 455.870 which is quite a substantial range.

The conference interval for car_2 falls between 753.481 and 810.660, and the prediction interval ranges from 658.450 to 905.692. There is a notably difference in the intervals for the two cars. With car_2 being newer, a more prestigious brand, more horsepower, and lower mileage, this should yield a substantial difference in the predictions if the assumption behind our model is true.

Looking at the statistics for the price interval, 50% of the car prices fall between this range 429.850 and 589.000 which gives a difference of 159.150, a much narrower range than our predicted intervals.

Min.	1st Qu	Median	Mean	3rd Qu	Max
199000	429850	499000	518588	589000	1599900

Figure 23, table with statistics for the price variable

The question one could pose is whether these prediction intervals are satisfactory, or if they are too wide. Further model testing, inclusion of more variables, transformation of variables or testing non-linear relationships in models would most likely narrow the intervals. But this is beyond the scope of this study.

5 Conclusions

Our initial goal was to find a linear multiple regression model that could predict the selling price of used electric cars on Blocket, as well as provide insights into the statistical inference of the chosen variables in the model on the prices. We posed the following questions:

1. What are the characteristics of our dataset, and how can we use it in a regression analysis?

The dataset covered a relatively narrow range, intentionally chosen to avoid an imbalanced data set. Despite this approach, the data did contain extreme values for the numeric variables, which is to be expected in a randomly selected dataset. Additionally, there was a problem with too few observations for some color and brand categories, necessitating the grouping of variables to stabilize the model. The distribution of the price variable was left-skewed, with very few observations above 800,000. Due to the limited number of high-priced cars, the model we found cannot be said to be representative of those price levels.

1. Which regression model provides the most accurate representation of the relationship between the variables and yields the best price predictions?

The final model was determined by removing high leverage points from the training set and conducting a best subset selection on the variables. The selected model included the following variables: mileage, horsepower, car_brand, and year_category.

On the test dataset, the model yielded an RMSE of 110.128 and an adjusted R^2 of 0.8116. This indicates that the variables included in the model collectively explain approximately 81% of the variance in the response variable.

It would have been desirable to have a smaller RMSE, since most cars did lie in a relatively narrow price range. Non-linear models could perhaps give more accurate predictions, but as stated in the introduction, they are also harder to interpret.

2. What is the relationship between the response variable price and the predictor variables' mileage, horsepower, car brand, color, and model year?

As anticipated, then the color variable had minimal impact on the prices and was consequently removed from the final model. While certain brands within the car brand category significantly influenced prices, others did not. Working with this variable was challenging due to the limited representation of some brands, necessitating their grouping into a more diverse category labeled "Other." This variable primarily provided insights into brand effects, which are more nuanced compared to, for instance, a car model variable. Although information on car models was collected, it was more uneven distributed than the car brand variable and was therefore omitted from the analysis.

The mileage variable was expected to exhibit a strong negative linear relationship with the price category, but this was not the case. While it did exert a significant influence on prices, it was not as pronounced as the horsepower variable, which displayed the strongest linear relationship with the price category. This was indicated by the best subset selection where horsepower was the first variable to be included, and as well seen in the correlation matrix with the numeric variables.

One hypothesis is that the mileage variable may play a more significant role in price determination for fossil-driven cars compared to electric cars, which would be an interesting finding if validated.

Additionally, the model year category also influenced price settings, with newer cars yielding higher prices. However, the variable had very few observations before 2020, prompting the creation of a category for "old" cars. Although negatively correlated with the mileage variable, it had an acceptable VIF-score, so both variables were obtained in the model.

5.1 Reflections

The final model chosen exhibited some issues in the diagnostic plots. The transformation of the response variable performed in model 2 showed a slight improvement, particularly in addressing heteroscedasticity and normalizing the distribution of the residuals. This suggests that the relationship between the response and predictor variables in our dataset may be better described by a non-linear model, which could potentially lead to better predictions. Instead of transforming the response variable, another option would have been to explore transformations of the predictor variables.

It can be debated whether removing the high leverage points from the training set was the appropriate decision. Removing observations should always be approached with caution, as they are integral to the dataset, and their removal may lead to a less representative model. In this study, they were removed to improve the model, despite potential concerns regarding the representativeness of the sample. To address these issues, it would have been beneficial to have more data points covering extreme values in the variables and a greater number of observations in qualitative variables with few common values. This would have helped to mitigate the need for removing observations and improve the overall robustness and generalizability of the model.

Furthermore, there remains an irreducible error in our models that needs to be addressed. As previously mentioned, there are likely other factors driving the prices of electric cars compared to fossil-driven cars, such as battery specifications and charging time. A subsequent study comparing the influencing factors in fossil-fuel cars and electric cars would be of interest. Such a study could reveal differences in how the different fuel types are presented on Blocket. Currently, all cars have the same information on the ads page but given the increasing prevalence of electric cars and potential differences in price-setting variables, it may be worth evaluating and updating this approach.

Reference List

Allison, Paul (2012, Sep). *When Can You Safely Ignore Multicollinearity?* Retrieved 25.04.2024 from https://statisticalhorizons.com/multicollinearity/ -

Bevans, R. (2023, June 22). *Simple linear regression | An easy introduction & examples*. Retrieved 19.04.2024 from Scribbr. https://www.scribbr.com/statistics/simple-linear-regression/

Bevans, R. (2023, June 22). *Multiple Linear Regression | A Quick Guide (Examples).* Retrieved 19.04.2024 from https://www.scribbr.com/statistics/multiple-linear-regression/ -

Cialdella, L (2020) *When do we log transform the response variable? Model assumptions, multiplicative combinations and log-linear models* from Casual Inference Data analysis and other apocrypha. Retrieved 27.04.2024 from https://lmc2179.github.io/posts/multiplicative.html -)

James, G. and Witten, D. Hastie, T. Tibshirani, R. (June 2023). *An Introduction to Statistical Learning with Applications in R. Second Edition*. Retrieved April 2024 from https://www.statlearning.com/ -

Theoretical Questions

1. **Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s, beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.**

A QQ plot-provides insights into whether values, such as observations or residuals, are drawn from a population with a normal distribution. If the observations on the QQ-plot align along a straight line, it indicates that the sample follows a normal distribution. QQ stands for quantile-quantile because we divide the normal distribution, typically the standard normal distribution, into quantiles based on the number of observations. This results in a plot where the observed values are plotted against the theoretical quantile values. If this assumption holds true, the plot will display a straight line.

Checking the residuals for normality is crucial when assessing a linear regression model since we assume that the residuals follow a normal distribution. The validity of this assumption can be verified using a QQ-plot.

2. **Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?**

Yes, in general, machine learning is oriented toward developing models capable of making accurate predictions on unseen data. This objective is inherent in the definition of machine learning, which aims to enable models to learn from data and perform predictive analyses. Examples of machine learning tasks include clustering, where models analyze data and partition it into clusters, as well as tasks like price prediction, image recognition, and speech recognition.

In contrast, statistics encompasses not only the development of predictive models but also the analysis of how variables influence these predictions, a field known as statistical inference. In statistics, the focus extends beyond solely achieving accurate predictions; it involves understanding the variables and drawing conclusions about why models perform well in prediction tasks. Examples of statistical tasks include predicting prices, as in this study, while simultaneously investigating factors that influence prices.

3. **Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?**

A prediction interval provides an estimate of the range within which the true values of Y are likely to deviate from the predicted value $\hat{Y}$. In contrast, while a prediction interval predicts a single point, a confidence interval surrounds the mean. While a prediction interval indicates where an individual observation point is likely to fall (95% of the times), a confidence interval indicates where 95% of the true average values of the dependent variable are expected to lie. Predicting an individual point entails greater uncertainty compared to predicting an average, resulting in prediction intervals being wider than confidence intervals.

4. **Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$. Hur tolkas beta parametrarna?**

$\beta 0$ represents the intercept of the multiple linear regression model, indicating the value of Y when all predictor variables are set to 0. The coefficients β preceding the x variables measure the magnitude of the influence each variable has on the response variable Y, while holding the other predictor variables constant. For instance, if $\beta 1 = 30$, implies that, when there is no change in the other variables, Y will increase by 30 units with a one-unit increase in x1.

ε represents the irreducible model error, reflecting the variability in Y that is not accounted for by the predictor variables. It is an essential component of the linear regression function, emphasizing that we cannot perfectly estimate Y based on the predictor variables alone. The irreducible error term may stem from unmeasured variables not included in the model or variations in the observation settings (James, p. 18ff).

5. **Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?**

The measurements BIC, AIC, adjusted R^2, or Cp can be used to evaluate models with a different number of variables. These metrics penalize the training error to provide estimates for the test error, making them suitable for variable selection as well as model selection.

However, when evaluating different models' predictive capabilities, it's recommended to either partition the dataset into train, validation, and test sets, or to use methods like cross-validation.

6. **Förklara algoritmen nedan för ”Best subset selection”**

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

The concept of Best Subset Selection involves testing all possible combinations of variables in models, as well as different numbers of variables in the models, and selecting the best one based on the smallest RSS (Residual Sum of Squares) or the largest R^2 value. The process begins by creating a null model with no predictors. Subsequently, all possible models are tested, starting with models containing only one predictor, followed by models with two predictors, and so forth. This exhaustive search is conducted on the training set to identify models with low training error.

To assess the models' ability to generalize to unseen data, they are evaluated on a validation set to estimate the prediction error. Metrics such as Cp, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), or adjusted R^2 are commonly used for this purpose.

When employing cross-validation, the process is extended by training the models on each training fold and validating them on the remaining folds. At the end of this process, the validation errors from each fold are averaged to determine the best model.

7. Ett citat från statistikern George Box är: "All models are wrong, some are useful." Förklara vad som menas med det citatet.

This implies that we will never achieve a perfect estimate for f(x). There exists a reducible error that we can minimize by refining the model and working with the data. However, there will always be an irreducible error that impacts the model, highlighting that the model is a mere representation of reality and not an exact depiction of it. By acknowledging that all models are flawed, this underscores the aspect of representing reality rather than presenting absolute truth. Despite this, some models can still be valuable because we can get insights from them by improving the model's estimation and reducing the reducible error within our models.

Data Collection - Group Assignment

I collaborated with Camilla, Sylvia, Magesh, and Bushra. During the initial days, we held several meetings to touch base and assess our progress. However, our group synergy was not optimal due to variance in activity levels among members. To improve collaboration, we could have clarified the preferred language for communication, as some members seemed more comfortable speaking English. Fortunately, we had a diplomatic member in the group who ensured that everyone was included in discussions before proceeding. Another group member took on the role of asking all the questions, while I took on the role of driving the work forward and answering the questions posed.

A positive aspect was the lack of pretense among group members; everyone contributed meaningfully without unnecessary disagreement or ego. It was evident that we shared a common goal: to gather sufficient data within a short timeframe.

My strengths lie in initiating and structuring processes, and I am comfortable taking on leadership roles when necessary. However, I prefer to step back and allow others to lead while I reflect. In this assignment, I felt compelled to take a leading role, which hindered my ability to reflect adequately on the assignment and the process. I need to find a balance between leadership and reflection. In hindsight, I would have preferred to be better prepared for the assignment and allocate more time for discussion and reflection at the outset, rather than focusing on completing the task of getting the data collected.

Self -Evaluation

Challenges

Every aspect of the assignment presented challenges, from group assignment to handling variables and structuring the project. I often found myself overwhelmed and lost in a sea of information while seeking assistance online. I approached the process like assembling a 1000-piece puzzle of a green lawn, tackling one piece at a time in the hope that the overall picture would emerge eventually.

Grade

While I am satisfied with my achievement, I also recognize that there is still much to learn in this field. Therefore, I refrain from suggesting a grade for this assignment.