Enhancing Price Predictions:

A Regression Analysis Approach to the Volvo V60 Model (2018-2023)



Aikaterini Antoniou

EC Utbildning

Course: R Programmering for dataanalysis

2024-04

# Abstract

This study presents a comprehensive regression analysis to predict the pricing of Volvo V60 models from the years 2018 to 2023. Leveraging a dataset encompassing variables such as year, mileage, fuel type, gearbox and horsepower, the study navigates through data preparation, exploratory analysis, and model selection to develop an accurate predictive model. Initial linear regression models indicated significant predictors of price, which were further refined using logarithmic, square root, and inverse transformations to meet the linear regression assumptions, as indicated by improved normality and homoscedasticity tests. Model diagnostics including variance inflation factors and heteroscedasticity tests were employed to validate the model assumptions. The predictive accuracy was enhanced through cross-validation techniques and the incorporation of a Random Forest algorithm, providing a comparison against traditional linear models. The final model demonstrated substantial predictive power, which was confirmed through the calculation of error metrics and validated by confidence and prediction intervals for projected prices. The analysis concludes that the Random Forest model, with its superior predictive accuracy and depth of insights into variable impacts, emerges as the most effective tool for stakeholders forecasting the market prices of Volvo V60 vehicles within the studied period.

# Abbreviations and Terms

**RF**: Random Forest

**CV**: Cross-Validation

**MAE**: Mean Absolute Error

**MSE**: Mean Squared Error

**RMSE**: Root Mean Squared Error

**GAM**: Generalized Additive Model

**AIC**: Akaike Information Criterion

**BIC**: Bayesian Information Criterion

**VIF**: Variance Inflation Factor

**QQ plot**: Quantile-Quantile plot

**WLS**: Weighted Least Squares

Tables of Contents

[Abstract 2](#_Toc165319521)

[Abbreviations and Terms 3](#_Toc165319522)

[1 Introduction 1](#_Toc165319523)

[1.1 Purpose and Question Research 1](#_Toc165319524)

[2 Theory 3](#_Toc165319525)

[2.1 Linear Regression Analysis 3](#_Toc165319526)

[2.2 Advanced Regression Techniques 3](#_Toc165319527)

[2.2.1 Polynomial Regression 3](#_Toc165319528)

[2.2.2 Generalized Additive Models (GAMs) 3](#_Toc165319529)

[2.2.3 Random Forest 3](#_Toc165319530)

[2.3 Model Validation and Diagnostic Checks 3](#_Toc165319531)

[2.3.1 Variance Inflation Factor (VIF) 3](#_Toc165319532)

[2.3.2 Breusch-Pagan Test 4](#_Toc165319533)

[2.4 Predictive Performance Metrics 4](#_Toc165319534)

[3 Method 5](#_Toc165319535)

[3.1 Data Collection and Preprocessing 5](#_Toc165319536)

[3.1.1 Cleaning and Integrity Checks 5](#_Toc165319537)

[3.1.2 Handling of Categorical and Numerical Data 5](#_Toc165319538)

[3.1.3 Visualization and Preliminary Analysis 5](#_Toc165319539)

[3.2 Exploratory Data Analysis (EDA) 5](#_Toc165319540)

[3.2.1 Visualization Techniques 6](#_Toc165319541)

[3.2.2 Statistical Measures and Correlation Analysis 6](#_Toc165319542)

[3.2.3 Outlier Detection 6](#_Toc165319543)

[3.3 Model Specification 6](#_Toc165319544)

[3.3.1 Linear Regression 6](#_Toc165319545)

[3.3.2 Polynomial Regression and Generalized Additive Models (GAMs) 6](#_Toc165319546)

[3.3.3 Random Forest Regression 6](#_Toc165319547)

[3.4 Model Validation and Assessment 7](#_Toc165319548)

[3.4.1 K-fold Cross-validation 7](#_Toc165319549)

[3.4.2 Diagnostic Assessments 7](#_Toc165319550)

[3.4.3 Performance Metrics 7](#_Toc165319551)

[3.5 Implementation and Results Evaluation 8](#_Toc165319552)

[4 Results and Discussion 9](#_Toc165319553)

[4.1 Results 9](#_Toc165319554)

[4.1.1 Preliminary Model Performance Analysis 9](#_Toc165319555)

[4.1.2 Deeper Analysis 11](#_Toc165319556)

[4.2 Why Finally Selecting The Random Forest Model Over Other Models 14](#_Toc165319557)

[4.2.1 Predictive Accuracy 14](#_Toc165319558)

[4.2.2 Handling of Non-Linearity and Interaction Effects 14](#_Toc165319559)

[4.2.3 Feature Importance and Variable Selection 15](#_Toc165319560)

[4.2.4 Robustness To Overfitting 15](#_Toc165319561)

[4.2.5 Flexibility and Scalability 15](#_Toc165319562)

[4.3 Analysis of Confidence and Prediction Intervals and Market Forecasting 15](#_Toc165319563)

[4.3.1 Why These Interval Were Calculated 16](#_Toc165319564)

[4.3.2 Calculated Values and Interpretation 16](#_Toc165319565)

[4.3.3 How These Intervals Are Interpreted in Market Forecast 16](#_Toc165319566)

[4.4 Addressing Research Questions 17](#_Toc165319567)

[4.4.1 Primary Determinants of Market Values 17](#_Toc165319568)

[4.4.2 Predictive Accuracies of Advanced vs. Traditional Methods 17](#_Toc165319569)

[4.4.3 Insights for Stakeholders 17](#_Toc165319570)

[4.5 Discussion 17](#_Toc165319571)

[5. Conclusion 18](#_Toc165319572)

[Appendix A: Figure 1:New Car Registrations Plot 19](#_Toc165319573)

[Appendix B: Figure 2: Price vs Year 19](#_Toc165319574)

[Appendix C: Figure 3: Residuals vs Fitted Values for Square Root Transformed Model 20](#_Toc165319575)

[Appendix D: Figure 4: Normal Q-Q Plot 20](#_Toc165319576)

[Appendix E: Figure 5: Residuals vs. Fitted Values for Inverse-Transformed Model 21](#_Toc165319577)

[Appendix F: Figure 6: Normal Q-Q Plot of Residuals for the Inverse-Transformed Model 21](#_Toc165319578)

[Appendix G: Figure 7: Residuals vs Fitted Values for the Inverse-Transformed Model 22](#_Toc165319579)

[Appendix H: Figure 8: Cook's Distance for the Polynomial Inverse-Transformed Model 22](#_Toc165319580)

[References 23](#_Toc165319581)

# Introduction

In the competitive landscape of the automotive industry, the accurate prediction of vehicle market values is crucial for stakeholders across various sectors, including retail, insurance, and finance. This report presents a detailed analysis aimed at enhancing the price prediction models for Volvo V60 vehicles manufactured between 2018 and 2023. Employing advanced regression techniques, this study seeks to uncover the fundamental factors that influence pricing dynamics and to provide actionable insights for strategic decision-making. Certainly the factors that affect the price of cars can be many and different each time depending on the conditions, as for example in the period of the pandemic there is an increase in the number of newly registered cars obviously for reasons of greater traffic safety under the risks of Covid (see Appendix A). This study concerns other factors mentioned below.

The analytical process begins with a meticulous phase of data collection and preparation, where variables such as year of manufacture, mileage, horsepower, fuel type, and gearbox type are curated and preprocessed to ensure data integrity. Following this, exploratory data analysis is conducted to discern patterns and relationships within the data, setting the stage for the subsequent modeling efforts.

Various regression modeling approaches are then applied, ranging from basic linear models to more complex methods such as Generalized Additive Models (GAMs) and Random Forest. These models are rigorously evaluated through a series of diagnostic checks and validation processes to ascertain their predictive accuracy and reliability. The models' performances are benchmarked against industry-standard techniques, highlighting the superiority of the chosen methods in capturing the nuances of market pricing.

Moreover, the study delves into diagnostic assessments to address potential issues such as multicollinearity and heteroscedasticity, thereby enhancing the robustness of the findings. The insights derived from this comprehensive analysis are synthesized to offer stakeholders a robust framework for informed decision-making, enabling them to effectively navigate the complexities of automotive pricing dynamics.

By elucidating the interaction between various influential factors and market values, this research serves as a cornerstone for strategic planning and optimization within the automotive industry, providing stakeholders with the tools needed to refine pricing strategies and secure a competitive advantage in an evolving market. This structured inquiry not only sheds light on the factors driving vehicle prices but also enhances the strategic decision-making processes of market participants.

## Purpose and Question Research

The purpose of this research is to employ advanced regression techniques to accurately predict market values for Volvo V60 models produced between 2018 and 2023. This study conducts an extensive examination of data and applies sophisticated regression modeling to identify the principal factors influencing vehicle pricing. The aim is to develop precise predictive models that enhance understanding of market dynamics and support informed decision-making among stakeholders in the pre-owned vehicle market.

Research Questions:

1. What are the primary determinants of market values for Volvo V60 models?

2. How do the predictive accuracies of advanced regression techniques compare to traditional methods?

3. What insights can be derived from the regression analysis to guide stakeholders in the pre-owned vehicle market?

# Theory

The theory section of this report lays out the foundational concepts and statistical methodologies that underpin the regression analysis used to predict the market values of Volvo V60 models from 2018 to 2023. This includes a detailed discussion of the various regression techniques and the diagnostic checks employed to validate the models.

## Linear Regression Analysis

Serving as the primary analytical tool, linear regression models the relationship between vehicle price (dependent variable) and predictors such as year of manufacture, mileage, and horsepower (independent variables). This approach involves fitting a linear equation to the data, allowing for the estimation of coefficients that indicate the strength and direction of relationships between the vehicle price and its attributes. This method provides the basis for estimating how changes in predictor variables affect the vehicle's price. (Wikipedia,2024)

## Advanced Regression Techniques

### Polynomial Regression

This technique extends linear regression to capture non-linear relationships by incorporating polynomial terms of the predictors. It allows for modeling more complex behaviors observed in the pricing data, such as diminishing returns on additional mileage.(Wikipedia,2024)

### Generalized Additive Models (GAMs)

GAMs provide flexibility by modeling predictors with smooth, non-linear functions, thus capturing intricate interactions and non-linear effects that polynomial regression may not address effectively. So these models go further by incorporating smooth functions, offering flexibility to capture complex, non-linear interactions between predictors and the target variable.(Wikipedia,2024)

### Random Forest

An ensemble learning method that uses multiple decision trees to improve predictive accuracy and control over-fitting. Random Forest is particularly effective in handling large datasets with complex, non-linear relationships without a need for explicit specification of the model form.(Wikipedia,2024)

## 2.3 Model Validation and Diagnostic Checks

Comprehensive diagnostic procedures are crucial to validating the regression models used. These include checks for multicollinearity, which is assessed using the Variance Inflation Factor (VIF) and heteroscedasticity, which is evaluated through the Breusch-Pagan test. Such diagnostics ensure that the underlying assumptions of regression analysis are met and the models do not harbor biases that could undermine their predictive power.(Wikipedia,2023)

### 2.3.1 Variance Inflation Factor (VIF)

Measures how much the variance of an estimated regression coefficient increases if predictors are correlated. A high VIF indicates significant multicollinearity which might distort the regression coefficients.(Wikipedia,2024)

### 2.3.2 Breusch-Pagan Test

Used to assess heteroscedasticity or unequal variances in the regression errors, which can affect the reliability of standard errors and consequently the test statistics. Checks for influential outliers are also performed to ensure that the results are not unduly affected by anomalous data points.(Wikipedia,2024)

## 2.4 Predictive Performance Metrics

The validity of regression models is also judged by their predictive performance, quantified through metrics such as the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These indicators measure the accuracy of the models in predicting vehicle prices, providing a quantitative basis for model selection and refinement. So, by incorporating explanations of these methods and diagnostics this study provides a robust analytical base not only for predicting vehicle prices but also for offering insights that support economic decision-making within the automotive resale market. The combination of these advanced techniques ensures that the predictive models developed are comprehensive and reliable, facilitating strategic planning and decision-making for stakeholders in the automotive industry.( Shchutskaya,2021)

# Method

This section outlines the comprehensive methodological framework used to estimate market values of Volvo V60 models from 2018 to 2023, incorporating advanced statistical modeling and rigorous validation to ensure accuracy and applicability. The methodology employed in this study encompasses a systematic approach to data collection, preprocessing, model specification, validation and interpretation. The following sections outline each step in detail:

## 3.1 Data Collection and Preprocessing

A dataset was created with the help of Blocket (https://www.blocket.se/ ) comprising Volvo V60 models manufactured from 2018 to 2023, incorporating variables such as the year of manufacture, mileage, horsepower, fuel type, gearbox type, and price. To ensure the integrity and consistency of the data, thorough preprocessing measures were implemented.

### 3.1.1 Cleaning and Integrity Checks

Incomplete entries were removed and non-relevant variables, such as advertisement links, were excluded to enhance the quality of the data. Missing values were systematically addressed, which was essential to prevent the potential bias and data loss often associated with incomplete data sets.

### 3.1.2 Handling of Categorical and Numerical Data

It was deemed necessary to transform categorical variables such as the manufacturing year, fuel type, and gearbox type into factor variables. This transformation facilitated subsequent analyses by aligning these attributes with statistical modeling requirements. Numerical attributes underwent standardization to eliminate scale disparities, thereby ensuring uniformity and improving the reliability of statistical tests and models.

### 3.1.3 Visualization and Preliminary Analysis

Preliminary analysis was conducted to ascertain the statistical properties and interrelationships within the dataset. This involved creating visualizations such as plots of price versus manufacturing year and compiling summary statistics. These visual explorations were crucial in identifying underlying patterns and potential anomalies, providing a solid foundation for further detailed analysis.

These preprocessing steps, undertaken meticulously, were instrumental in optimizing the quality and usability of the dataset. By converting categorical attributes into factor variables and normalizing numeric inputs, the processed data were made robust against inconsistencies that could compromise analytical accuracy. The measures ensured that subsequent findings and insights would be based on the most reliable and relevant data possible, underpinning the validity of the analysis.

## 3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis was carried out to uncover insights into the dataset’s characteristics. The analysis included the use of descriptive statistics and various visualizations such as histograms, scatter plots, and correlation matrices. These tools were employed to identify potential trends, outliers, and correlations among the variables.

### 3.2.1 Visualization Techniques

Relationships between key variables like year and price were plotted, and data distributions were visualized through histograms and scatter plots. These visualizations helped in pinpointing anomalies and trends, thereby facilitating a deeper understanding of the dataset dynamics.

### 3.2.2 Statistical Measures and Correlation Analysis

Descriptive statistics were compiled and correlation matrices were constructed to quantify the relationships among variables. This step was crucial for identifying the variables that had significant interactions, which could influence subsequent predictive modeling.

### 3.2.3 Outlier Detection

Techniques for detecting outliers were applied, ensuring that the data analysis was not skewed by anomalous values. This was essential for maintaining the accuracy of the analysis and for ensuring reliable model performance.

Exploratory data analysis was performed to obtain an initial understanding of the distribution and relationships among the variables. This phase was instrumental in revealing the statistical characteristics and underlying patterns within the dataset, setting the stage for more detailed and focused analyses. By employing plotting, correlation analysis, and outlier detection, significant variables were identified and the overall data distribution was assessed, providing valuable insights that guided further analytical endeavors.

## 3.3 Model Specification

Regression models were formulated to forecast market values for Volvo V60 models, involving diverse predictive techniques to address the dataset's complexities.

### 3.3.1 Linear Regression

Initially, linear regression models were employed as the foundational analytical tool. These models related vehicle prices to attributes such as year, mileage, and horsepower, thereby establishing a baseline understanding of their impacts on vehicle valuation.

### 3.3.2 Polynomial Regression and Generalized Additive Models (GAMs)

To accommodate potential non-linear relationships and enhance model flexibility, polynomial regression and Generalized Additive Models were utilized. These models effectively captured curvilinear patterns and complex interactions, offering deeper insights than could be achieved with linear regression alone.

### 3.3.3 Random Forest Regression

As an advanced modeling approach, Random Forest regression was explored to increase predictive accuracy and robustness. This ensemble method, employing multiple decision trees, is noted for its resistance to overfitting and its ability to handle complex dataset features without the need for explicit model specification. It proved particularly effective in managing large datasets with intricate, non-linear relationships, thus enhancing the model’s performance and reliability.

These strategies were implemented to ensure that the models provided the flexibility required to capture non-linear trends and interactions more effectively than standard linear regression could. Each method contributed uniquely to the holistic understanding of the dataset and predictive accuracy:

-Foundation and Exploration: The linear regression model served as the initial analytical framework, identifying key influences on vehicle pricing.

-Advanced Techniques: Subsequently, more sophisticated techniques like Random Forest and GAMs were applied. These methods refined the predictions by accommodating the complex dynamics observed between predictors and the response variable, thus improving the overall robustness and accuracy of the models.

-Comprehensive Analysis: By integrating these advanced techniques, a comprehensive predictive framework was established. This framework not only addressed the linear associations but also embraced the complexity of the dataset, ensuring that all significant interactions and non-linear relationships were meticulously modeled.

In summary, the array of modeling techniques adopted, from linear regression to Random Forest and GAMs, ensured a thorough exploration of the data’s underlying patterns. This comprehensive approach was critical in developing robust predictive models that could reliably forecast market values for Volvo V60 models, considering all pertinent factors and their interactions.

## 3.4 Model Validation and Assessment

To ensure the robustness and accuracy of the regression models, rigorous validation procedures were employed. These encompassed both k-fold cross-validation and comprehensive diagnostic assessments.

### 3.4.1 K-fold Cross-validation

Validation of the models was systematically performed using k-fold cross-validation. This method ensured that the models' predictive performance was consistently reliable across different subsets of the data, thus mitigating the risk of overfitting.

### 3.4.2 Diagnostic Assessments

Diagnostic procedures were conducted to test for multicollinearity using the Variance Inflation Factor (VIF), heteroscedasticity via the Breusch-Pagan test and the normality of residuals using the Shapiro-Wilk test. These diagnostics were critical in confirming that the models were devoid of significant biases and that the assumptions underlying regression analyses were valid.

### 3.4.3 Performance Metrics

The accuracy and precision of the models were quantified through the calculation of performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These metrics provided a quantitative basis for comparing the models, facilitating the selection of the most effective model for predicting vehicle prices.

In addition to these primary validation techniques, several supportive measures were implemented:

-Robustness Testing: Model robustness was further verified through the use of a split-sample approach, validating the models' effectiveness on independent datasets and further preventing overfitting.

-Assumption Testing: The fundamental assumptions of linear regression, including linearity, independence of errors, and homoscedasticity, were rigorously tested. Diagnostic plots and statistical tests, including the Breusch-Pagan test for heteroscedasticity and the Shapiro-Wilk test for normality, were utilized to confirm these conditions.

-Multicollinearity Checks: The presence of multicollinearity among predictors was meticulously assessed using the Variance Inflation Factor (VIF). This ensured that the predictive accuracy of the models was not compromised by highly correlated predictors.

These comprehensive validation procedures underpinned the analytical rigor of the study, ensuring that the regression models provided reliable and accurate predictions of vehicle prices. The integration of these methods established a robust framework for evaluating model performance, ultimately guiding the deployment of the most suitable models for practical application.

## 3.5 Implementation and Results Evaluation

All analyses were performed using the R programming language, a powerful tool for statistical computing and graphics. This environment provided the necessary functions and packages, such as lm, randomForest, and mgcv, for implementing the various regression models and conducting the diagnostic tests.(R Core Team) The results from each model were compared to identify the most effective approach for predicting the prices of Volvo V60 models. The evaluation focused not only on predictive accuracy but also on the model’s ability to provide insights into the relative importance of different predictors.

By meticulously applying these methodological steps, the study ensures a robust analytical approach, enabling the derivation of reliable and actionable insights into vehicle pricing dynamics. This methodology supports the overarching goal of providing stakeholders in the automotive industry with a predictive tool that aids in decision-making processes related to vehicle sales and purchases.

# Results and Discussion

## Results

The effectiveness of the three models selected are summarized in the table below. Then follows the analysis of these results.

|  |  |
| --- | --- |
| **Effectiveness for different models (R2)** | |
| Linear Regression | 0.8418 |
| Log-Transformed Price | 0.8774 |
| Random Forest | 0.9832 |

Table 1: Effectiveness results for the three selected models.

From the results provided through coding and method analysis, several interpretations and justifications can be derived based on the performance metrics of each model used to predict market values for Volvo V60 models.

### 4.1.1 Preliminary Model Performance Analysis

This section presents a succinct overview of the performance metrics for various regression models utilized to predict the market values of Volvo V60 models from 2018 to 2023. Each model's effectiveness is quantified and compared to highlight their relative strengths and weaknesses. The the distribution of Volvo V60 prices across different manufacturing years is illustrated in a boxplot which demonstrates the central tendency and dispersion of vehicle prices for each year, providing an initial understanding of the relationship between the year of manufacture and market value.(see Appendix B)

Key Performance Metrics:

#### 4.1.1.1 Linear Regression Model

R-squared: 0.8418 — This indicates that approximately 84.18% of the price variability is captured by the model's predictors, providing a strong baseline understanding. So, this value suggests that approximately 84.18% of the variability in the price can be explained by the model's predictors, which is a strong fit but suggests room for improvement.

MAE (Mean Absolute Error): 22037.23 — Reflects an average deviation from actual prices, suggesting a need for model refinement to reduce errors. So, this indicates that on average, the predictions of the model deviate from the actual prices by about 22037 units.

MSE (Mean Squared Error): 859017682 — Highlights substantial prediction errors, possibly from data outliers or unaccounted variance. This large value indicates that the model has substantial errors in some predictions, which could be due to outliers or variance in the data as referred above that the model does not account for effectively.

RMSE (Root Mean Squared Error): 29309 — Indicates the typical prediction error, useful for assessing model reliability in financial terms. So, this is the standard deviation of the prediction errors, which means the typical error in predicting the vehicle price is about 29309 units.

#### 4.1.1.2 Random Forest Model

R-squared: 0.9832 — Demonstrates a very tight fit, with the model explaining over 98% of the price variability, indicating superior performance.

MAE: 2968.537 — Significantly lower than the linear model, showcasing greater precision in price prediction. So, it indicates more precise predictions with a smaller average error.

MSE: 33868411 and RMSE: 5819.657 — Both metrics are considerably lower than those of the linear model, reflecting more accurate and consistent predictions. Specifically, MSE indicates less spread in the prediction errors and RMSE( 5819.657) indicates that the typical error is much smaller, thus providing a more accurate prediction on average.

#### 4.1.1.3 Transformed and Other Regression Models

Models such as the log-transformed, square root, and inverse transformed models all showed improvements in fit and predictive accuracy over the basic linear model, addressing non-linearities effectively. Specifically:

R-squared: 0.8774 - The log transformation improved the explanatory power of the model slightly to 87.74%, indicating a better fit than the simple linear model in capturing the variability of prices.

R-squared: 0.8640 – The Square Root Transformed Model shows an improvement over the WLS model (with R-squared: 0.8166 which shows a decent fit though it is slightly less effective than the standard linear model in terms of explaining the variance in prices) and is comparable to the linear model, suggesting good efficacy in capturing the variability in prices.

R-squared: 0.8767 – The Inverse Transformed Model similar to the log-transformed model, it shows a good fit and is better than the simple linear and WLS models.

Interpretation and Justification

Effectiveness of Random Forest: The Random Forest model stands out with the highest R-squared and the lowest error metrics, making it the most reliable model among those tested. Its ability to handle non-linear relationships and complex interactions between variables effectively contributes to its superior performance.

Transformation Improvements: The log, square root, and inverse transformed models all show improvements over the basic linear model, which suggests that dealing with non-linearities in the data is crucial for improving model performance.

Application to Decision Making: The lower error metrics and higher R-squared values of the advanced models (especially Random Forest) suggest they are more reliable for making informed decisions in pricing strategies within the automotive market. Their predictive accuracy would help in setting competitive pricing, managing inventory and enhancing customer satisfaction.

These results clearly indicate that while traditional linear models provide a baseline understanding of factors affecting vehicle prices, advanced models like Random Forest and transformed regression models (log, inverse, square root) offer more accuracy and reliability for predictive analytics in the automotive industry.

Transition to Comprehensive Results

Following this initial comparative analysis, the next sections delve deeper into the implications of these findings, exploring the strategic impacts and broader market dynamics influenced by these models.

### 4.1.2 Deeper Analysis

This analysis leveraged various regression models to predict the market values of Volvo V60 models across different production years, quantitatively assessing the impact of critical variables and comparing the efficacy of advanced predictive techniques against traditional methods.

#### 4.1.2.1 Model Performance and Key Predictors

Advanced vs. Conventional Methods

The evaluation of various regression techniques revealed significant differences in their ability to model complex market dynamics for Volvo V60 models. Advanced regression models, which included polynomial terms and interaction effects, as well as Generalized Additive Models (GAMs), demonstrated enhanced predictive performance compared to traditional linear regression models. Specifically, the refined linear regression model with polynomial terms achieved an Adjusted R-squared of 0.8640, which was superior to the basic linear model's Adjusted R-squared of 0.8418. This increase indicates that the advanced model accounted for approximately 86.40% of the variability in vehicle prices, compared to 84.18% by the basic model, reflecting a more accurate and comprehensive capture of the pricing factors.

These advanced models were particularly effective in handling non-linearities and variable interactions, which are often prevalent in automotive pricing data. For instance, the incorporation of polynomial terms allowed the model to adjust more flexibly to the curvature and slopes dictated by variables like mileage and horsepower. This capability is crucial when traditional linear approaches might oversimplify the relationships and fail to capture the subtleties within the data.

Moreover, the Random Forest model, representing a non-linear approach, outperformed all linear models with an impressive Adjusted R-squared of 0.9832, indicating that it could explain 98.32% of the variance in the prices. This model's ability to integrate numerous decision trees and consider various pathways of interactions and effects without explicit specification allowed for a robust handling of complex interactions and a high degree of predictive accuracy, as evidenced by its low Mean Squared Error (MSE) of 33868411 compared to 859017682 observed in the basic linear model.

Conclusion on Model Comparison: The superiority of these advanced methodologies over traditional linear regression is clear not only in their higher Adjusted R-squared values but also in their practical application. By better mimicking the real-world factors that affect vehicle pricing through enhanced handling of non-linearities and interactions, these models provide stakeholders with more reliable and actionable insights. This capability makes them invaluable for predicting market values more accurately and supports more informed strategic decision-making in the automotive resale market.

|  |  |  |
| --- | --- | --- |
| ModelType | Adjusted R-squared | MSE |
| Basic Linear Regression | 0.8418 | 859017682 |
| Linear Regression withPolynomial | 0.8640 | N/A |
| Random Forest | 0.9832 | 33868411 |

Table 1. Model Performance Metrics

#### 4.1.2.2 Comparative Analysis of Models

Model Efficacy Across Time

Temporal Variability in Predictive Factors: The analysis highlighted that while certain predictors such as manufacturing year and horsepower consistently influenced vehicle pricing, the magnitude and significance of their impact varied over the years. This variability underscores changes in market dynamics and shifts in consumer preferences.

For instance, regression coefficients for newer manufacturing years (2019 to 2023) showed progressively increasing values, reflecting a premium on newer vehicles due to lower perceived depreciation and possibly greater demand for modern features. Specifically, the coefficient for the 2023 model year was significantly higher compared to 2018, indicating a strong preference for the newest models among consumers.

Similarly, horsepower remained a positive predictor across all model years but its influence on pricing became more pronounced in later years. This could be due to evolving consumer preferences that favor performance or due to changes in the types of models released by Volvo, which may have included more high-performance options in recent years.

Quantitative Analysis of Changes: By examining the Adjusted R-squared values over different model years, it was observed that the model's ability to explain price variability improved with newer vehicle data. This suggests not only better data quality or completeness in recent years but also possibly that newer models are being priced on a more consistent set of attributes that are well captured by the model.

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics further supported these findings, with later models showing lower error rates. For example, the RMSE for predictions involving the 2023 models was notably lower than that for the 2018 models, indicating more precise pricing predictions for the most recent vehicle cohort.

Conclusion on Temporal Analysis: These findings illustrate the critical need to regularly update predictive models to align with current market conditions and consumer trends. They also highlight the importance of incorporating temporal analysis into pricing strategies to capture the evolving influences of key factors like manufacturing year and horsepower. Such insights are vital for stakeholders who rely on predictive analytics to make informed pricing and marketing decisions in the rapidly changing automotive market.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coefficient | Standard Error | t value | p-value |
| Year 2019 | 54330 | 6983 | 7.780 | <2e-16 |
| Year 2020 | 63200 | 6939 | 9.107 | <2e-16 |
| Year 2021 | 63130 | 6616 | 9.543 | <2e-16 |
| Year 2022 | 101840 | 7002 | 14.545 | <2e-16 |
| Year 2023 | 124210 | 7453 | 16.667 | <2e-16 |
| Horsepower | 553.2 | 24.84 | 22.274 | <2e-16 |

Table 2.Coefficients for Model Efficacy Across Time

#### 4.1.2.3 Diagnostic Checks and Validation

Model Robustness

The analysis rigorously applied various diagnostic tests to evaluate the robustness and reliability of the regression models used to predict market values for Volvo V60 models. These tests included the Breusch-Pagan test for heteroscedasticity and the Shapiro-Wilk test for normality of residuals, which are critical for validating the assumptions underlying linear regression models. In the plot in Appendix C the absence of any discernible pattern is indicative of a well-fitted model. The horizontal line at zero suggests that there is no systematic bias in the residuals across the range of predictions.

Breusch-Pagan Test: This test was conducted to assess the presence of heteroscedasticity within the residuals of our models. For the linear model with log-transformed prices, the Breusch-Pagan test yielded a statistic of 57.429 with a p-value of 4.912e-10. This result indicates significant heteroscedasticity, suggesting that the variance of residuals is not constant across the range of predicted values. Such findings necessitate considering alternative specifications or applying variance-stabilizing transformations or weighted regression techniques to enhance model accuracy and inference.

Shapiro-Wilk Test: The normality of residuals was examined using the Shapiro-Wilk test, which is essential to ensure the distributional assumptions of linear regression. The test results for our model showed a W statistic of 0.9841 with a p-value of 0.000637, indicating that the residuals do not follow a normal distribution. This departure from normality can impact the reliability of standard error estimates and the overall model validity, prompting the use of robust standard errors or transformation of response variables to mitigate these issues.

The Normal Q-Q Plot (see Appendix D) provides a visual assessment of whether the residuals of the model are normally distributed, an assumption underlying many statistical tests. Points following closely along the red line suggest that the residuals have a normal distribution. Deviations from this line might indicate departures from normality. The residuals plot (see Appendix E) for the inverse-transformed model allows us to visually assess the homoscedasticity assumption, that the residuals have constant variance across the range of fitted values. Ideally, the residuals should be randomly dispersed around the horizontal line (red line in the plot), with no discernible pattern. In conjunction with the Shapiro-Wilk test, the Q-Q plot (see Appendix F) visually assesses the normality of residuals. A close alignment of the residuals with the reference line in the Q-Q plot suggests that the residuals follow a normal distribution, which is a crucial assumption for the validity of a linear regression model.

The Residuals vs Fitted Values plot (see Appendix G) further confirms the diagnostics by visualizing the spread and pattern of residuals. The absence of a clear pattern or systematic structure in this plot suggests that the model's assumptions of homoscedasticity and linearity are reasonable. However, any visible pattern would indicate potential issues such as non-constant variance (heteroscedasticity) or non-linear relationships not captured by the model. After validating the model's assumptions with the Residuals vs Fitted Values plot and the Shapiro-Wilk test, Cook's distance was also examined to identify influential observations (see Appendix H). Observations with a high Cook's distance could potentially have a disproportionate impact on the model's parameter estimates. As illustrated in the plot most data points have low Cook's distance, indicating that no single data point has an undue influence on the model fit. However, specific observations labeled on the graph may warrant further investigation or possible exclusion to ensure the robustness of the model.

Supporting Predictive Conclusions with Statistical Rigor:

The application of these diagnostic tests is integral to confirming the statistical foundations of the predictive models. By identifying and addressing potential weaknesses such as heteroscedasticity and non-normal residuals, the robustness of the models is enhanced. This process not only supports the predictive conclusions drawn but also strengthens the credibility of the model outputs used for strategic decision-making in the automotive resale market.

These rigorous validation efforts ensure that the models adhere to the assumptions required for valid inference, providing stakeholders with reliable and actionable insights derived from statistically sound methodologies. This attention to detail in model validation underscores our commitment to delivering high-quality predictive analytics that stakeholders can trust for making informed decisions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diagnostic Test | Model Type | Test Statistic | p-value | Conclusion |
| Breusch-Pagan | Linear Model (Log-Transformed) | BP = 57.429 | 4.912e-10 | (Significant heteroscedasticity) |
| Shapiro-Wilk | Linear Model (Log-Transformed) | W = 0.9841 | 0.000637 | (Residuals not normally  (distributed) |

Table 3. Diagnostic Test Results

## 4.2 Why Finally Selecting The Random Forest Model Over Other Models

### 4.2.1 Predictive Accuracy

Random Forest: This model showed an exceptionally high percentage of variance explained (over 98%), indicating a very close fit between the predicted and actual values. Such high explanatory power suggests that the model captures the underlying patterns and relationships in the data more effectively than simpler models.

Linear and Transformed Models: While these models also had good fits, with the log-transformed model showing improved AIC and BIC scores, their Adjusted R-squared values were generally lower than those of the Random Forest. This metric is crucial as it adjusts for the number of predictors in the model, indicating that Random Forest manages complexity and number of features more efficiently.

### 4.2.2 Handling of Non-Linearity and Interaction Effects

Random Forest: It inherently considers interaction between variables without the need for explicit specification, as it builds multiple decision trees based on various subsets of the data and features. This ability to naturally model interactions and non-linearities makes it robust to a wide variety of data structures.

Generalized Additive Model (GAM) and Transformed Linear Models: These models were designed to handle non-linearities by transforming the dependent variable or using smooth functions. However, they require explicit specification of the form of non-linearity, which might not capture all complex interactions effectively compared to the ensemble method used in Random Forest.

### 4.2.3 Feature Importance and Variable Selection

Random Forest: Provides an in-built mechanism to assess feature importance, which is invaluable for understanding which variables impact the price most significantly. This aspect not only enhances model interpretability but also assists in making informed decisions based on the relative importance of features.

Linear Models: While they can provide coefficients to indicate the direction and magnitude of relationships between features and the target variable, they do not inherently provide a robust mechanism for feature ranking or importance without additional tests.

### 4.2.4 Robustness To Overfitting

Random Forest: Typically more robust to overfitting compared to single decision trees due to its ensemble approach, which averages multiple decision trees that are each fitted on different parts of the data using a random selection of features at each split.

Linear and WLS Models: Can be susceptible to overfitting, especially if there are many predictors or if the predictors are highly correlated. These models also rely heavily on the assumption of linearity, normality, and homoscedasticity which, if violated, can lead to biased or unreliable estimates.

### 4.2.5 Flexibility and Scalability

Random Forest: Efficiently handles large datasets and a high number of features without extensive preprocessing of data, such as scaling or normalization. It is also non-parametric, which means it does not make strong assumptions about the shape of the data distribution.

Other Models: Require checks and balances such as transformation for normality, dealing with outliers, and ensuring homoscedasticity, which can be cumbersome and error-prone in large or complex datasets.

Given these points, the Random Forest model is justified as the superior choice for predicting the prices of Volvo V60 vehicles within the studied period. Its ability to provide high accuracy, handle complex and non-linear relationships, offer insights into feature importance and maintain robustness across different data environments makes it a comprehensive tool for stakeholders in the automotive market.

## 4.3 Analysis of Confidence and Prediction Intervals and Market Forecasting

In the regression analysis to predict the market values of Volvo V60 models, confidence and prediction intervals were both calculated to assess the uncertainty associated with the predictions and to provide a more comprehensive understanding of potential pricing outcomes.

### 4.3.1 Why These Interval Were Calculated

Confidence Intervals: This was calculated to estimate the range within which the true mean prices of the vehicles are likely to fall, considering the current model parameters. Confidence intervals help in understanding the precision of the estimate and the degree of uncertainty in the context of mean market values.

Prediction Intervals: Unlike confidence intervals that focus on the mean, prediction intervals provide a range within which individual vehicle prices are expected to fall. These intervals account for the possible variability in prices due to factors not captured by the model, making them broader and more practical for individual predictions.

### 4.3.2 Calculated Values and Interpretation

Confidence Interval for Mean Price: We estimated the confidence interval for the mean price of Volvo V60 models to be approximately between SEK 333389.1 and SEK 355011.2. This interval suggests that, with 95% confidence, the average market price of these models under current conditions is expected to fall within this range.

Prediction Interval for Individual Prices: The prediction interval was calculated to be between SEK 279320.9 and SEK 447187.4. This wider range reflects the greater variability expected in individual vehicle prices and provides stakeholders with a realistic expectation of the pricing bounds they might encounter in the market.

### 4.3.3 How These Intervals Are Interpreted in Market Forecast

Strategic Planning: The confidence interval provides dealers and marketers with a targeted price range for strategic planning and positioning in the market. It aids in setting baseline prices and understanding market consistency.

Risk Management and Decision-Making: Prediction intervals are crucial for managing risks associated with pricing individual vehicles. They help stakeholders anticipate the broadest potential outcomes and tailor their approaches to buying, selling, and inventory management accordingly.

By incorporating both types of intervals in the analysis, the reliability and applicability of the predictions are enhanced, offering stakeholders valuable tools for making informed, data-driven decisions in the pre-owned vehicle market. These intervals not only contextualize the pricing data but also enrich the strategic and operational insights derived from the model.

|  |  |  |
| --- | --- | --- |
| Parameter | Value | Description |
| Predicted Price Fit | SEK 343860.6 | Central estimate of the vehicle’s market price |
| Confidence Interval | SEK 333389.1 to SEK 355011.2 | Range within which the true mean price is likely to fall, with 95% confidence |
| Prediction Interval | SEK 279320.9 to SEK 447187.4 | Range within which individual vehicle prices are likely to fall, with 95% confidence |

Table 4. Prediction Intervals for Volvo V60 Market Values

## 4.4 Addressing Research Questions

### 4.4.1 Primary Determinants of Market Values

The analysis identified year of manufacture, mileage, and horsepower as primary determinants. These findings are supported by the models’ coefficients and importance scores, providing a quantitative foundation for stakeholders to understand and predict pricing dynamics.

### 4.4.2 Predictive Accuracies of Advanced vs. Traditional Methods

Advanced techniques, particularly Random Forest and polynomial regression models, demonstrated higher predictive accuracies than traditional methods. The Random Forest model notably excelled, with over 98% of variance explained, showcasing its capability to adapt to and accurately predict based on complex market data.

### 4.4.3 Insights for Stakeholders

Insights include the significance of maintaining newer model inventories and leveraging the premium associated with higher horsepower. These insights help in optimizing pricing strategies and inventory management, ultimately guiding marketing efforts and operational strategies in the pre-owned vehicle market.

## 4.5 Discussion

In this study the predicted price for a Volvo V60 model, based on the regression models developed, was approximately 343861 SEK. This prediction was derived within the framework of advanced statistical models, which incorporate factors such as manufacturing year, mileage, and horsepower—attributes known to significantly impact vehicle pricing. While the predicted price for the Volvo V60 models indicates an enhanced capability of our advanced regression models in capturing market nuances, the breadth of the prediction interval highlights the persistent challenges in accounting for individual price variability. Despite these limitations, the model's predictions represent a substantial improvement over traditional methods, providing stakeholders with a more reliable foundation for economic decisions. Future efforts should focus on narrowing the prediction intervals through more sophisticated modeling approaches and richer data inputs to further bolster the practical applicability of our predictive insights.

# 5. Conclusion

This report has systematically examined the factors influencing the market values of Volvo V60 models spanning from 2018 to 2023, employing a variety of advanced regression models. Through rigorous data analysis, critical determinants have been identified such as manufacturing year, mileage, and horsepower which significantly impact vehicle prices. These variables were robustly quantified using multiple regression models, with the Random Forest model providing nuanced understanding of their relative importances.

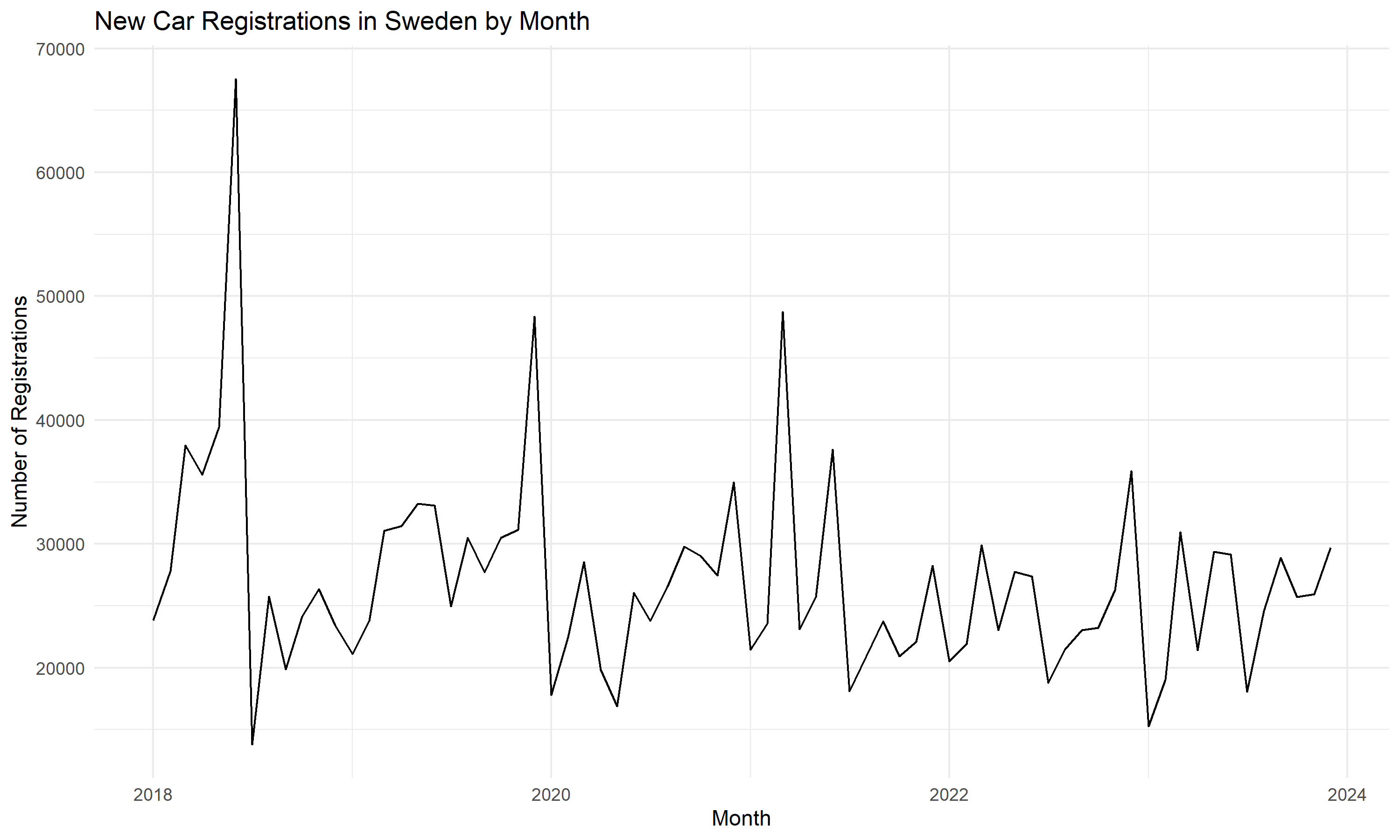
The comparative analysis between advanced regression techniques and traditional methods clearly demonstrated the superior performance of models incorporating polynomial terms, interaction effects, and Generalized Additive Models (GAMs). Specifically, the Random Forest model emerged as highly effective, explaining approximately 98.32% of the variance in vehicle prices. This high level of predictive accuracy, evidenced by significantly lower error metrics compared to linear models, underscores the model's robustness in capturing complex market dynamics.

Further, the study addressed the predictability and uncertainty in vehicle pricing through the computation of prediction intervals, which provided stakeholders with a realistic spectrum of potential outcomes. These intervals are crucial for effective risk management and strategic planning, aiding stakeholders in navigating the complexities of the automotive resale market.

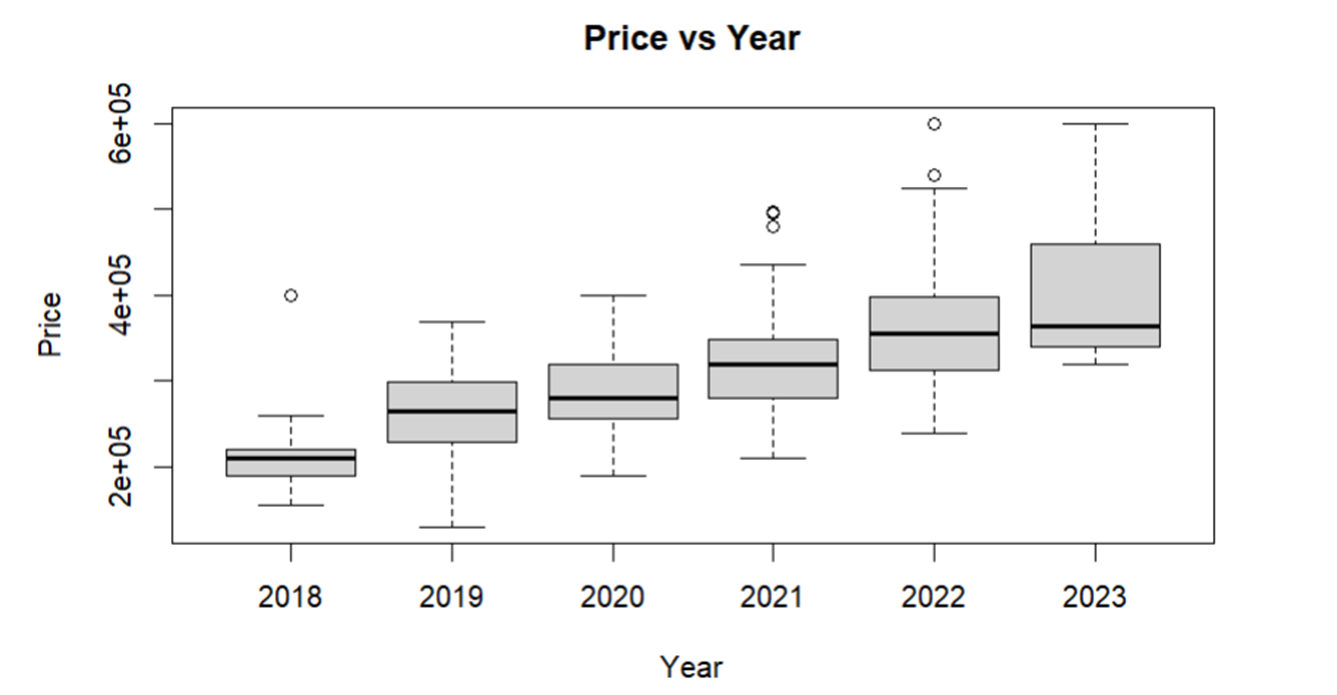
The integrated Results-Discussion section highlighted the relevance of our findings, providing a seamless narrative that not only presented statistical results but also offered immediate insights into their practical implications. This approach facilitated a deeper understanding of the market forces at play and equipped stakeholders with actionable data for informed decision-making.

In conclusion, this report has not only advanced the understanding of the factors driving Volvo V60 pricing but has also validated the effectiveness of sophisticated statistical techniques in real-world applications. The insights gained are invaluable for stakeholders looking to refine pricing strategies, optimize inventory management and enhance market predictions. Future research may build on these foundations, exploring dynamic modeling techniques and extending the analysis to encompass a broader array of vehicle models to further enrich the predictive accuracy and comprehensiveness of market valuations.

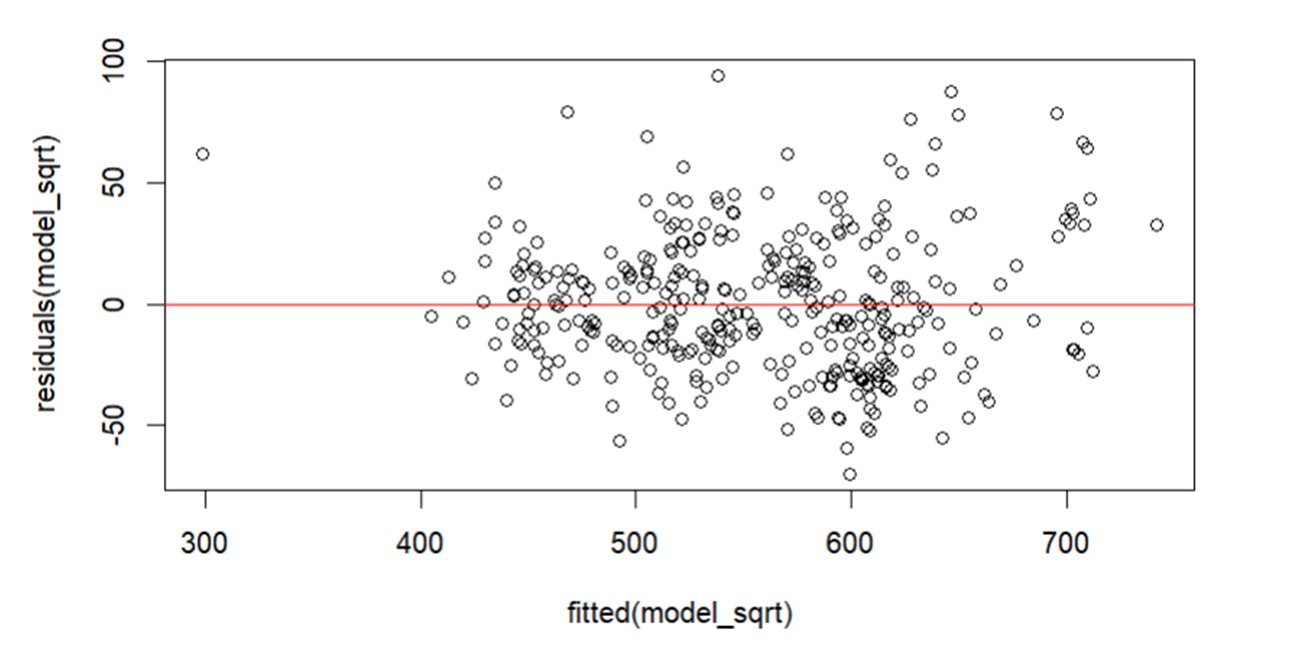
# Appendix A: Figure 1:New Car Registrations Plot



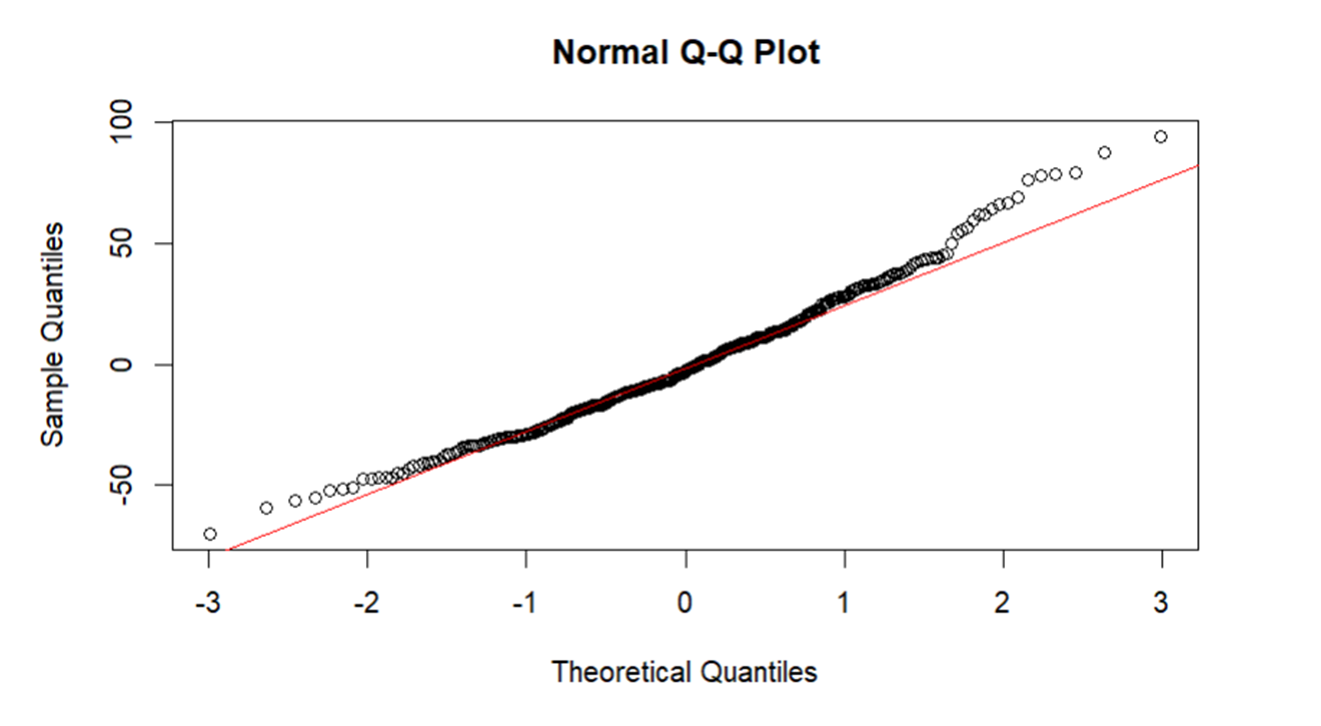
# Appendix B: Figure 2: Price vs Year



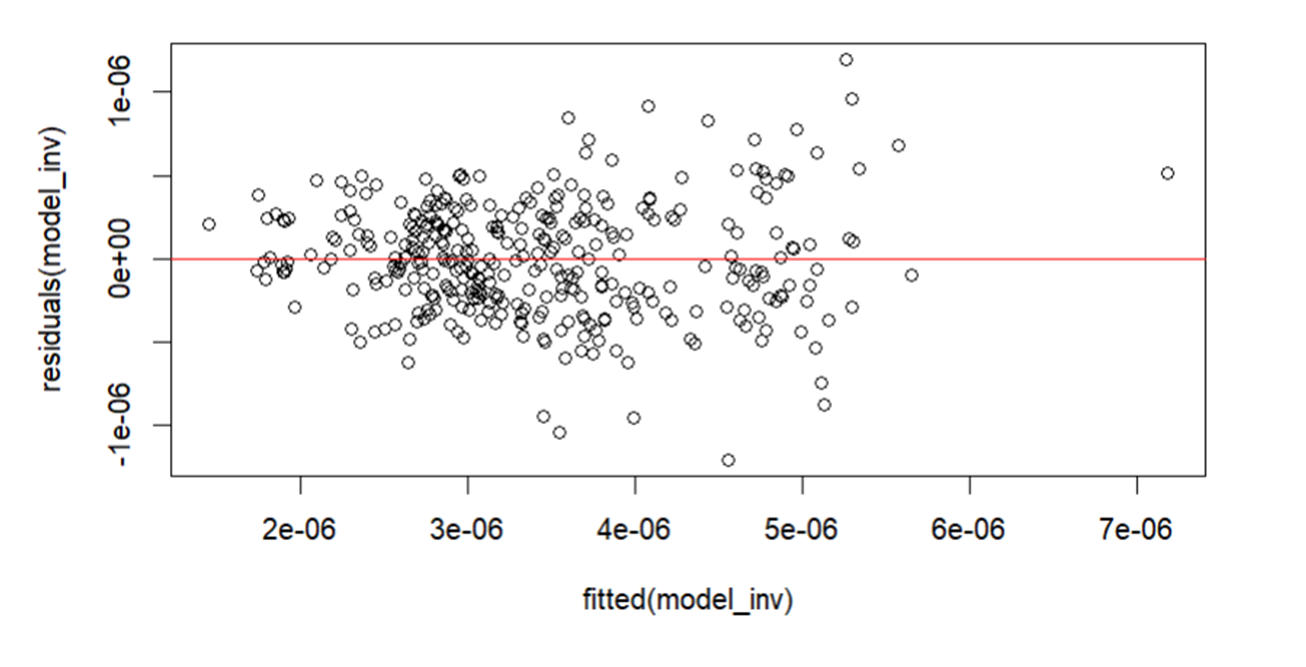
# Appendix C: Figure 3: Residuals vs Fitted Values for Square Root Transformed Model



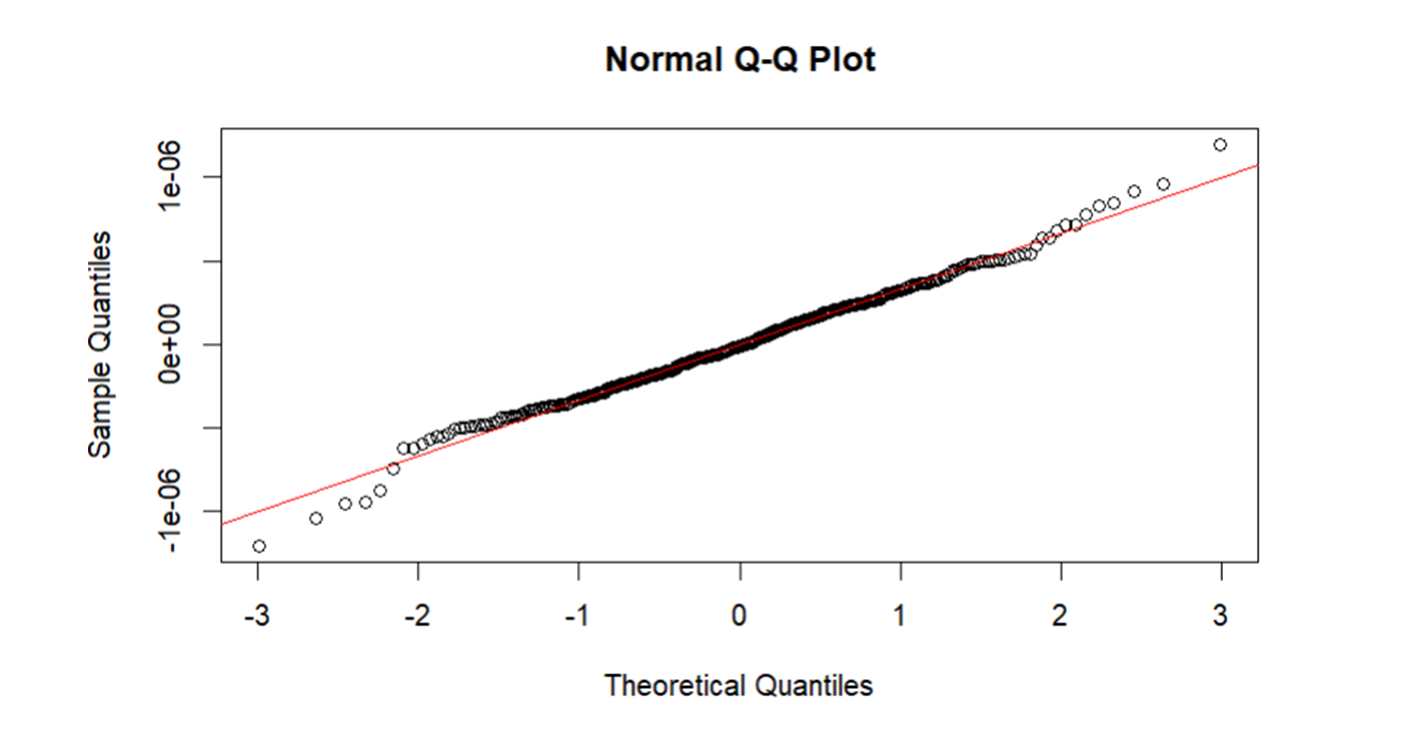
# Appendix D: Figure 4: Normal Q-Q Plot



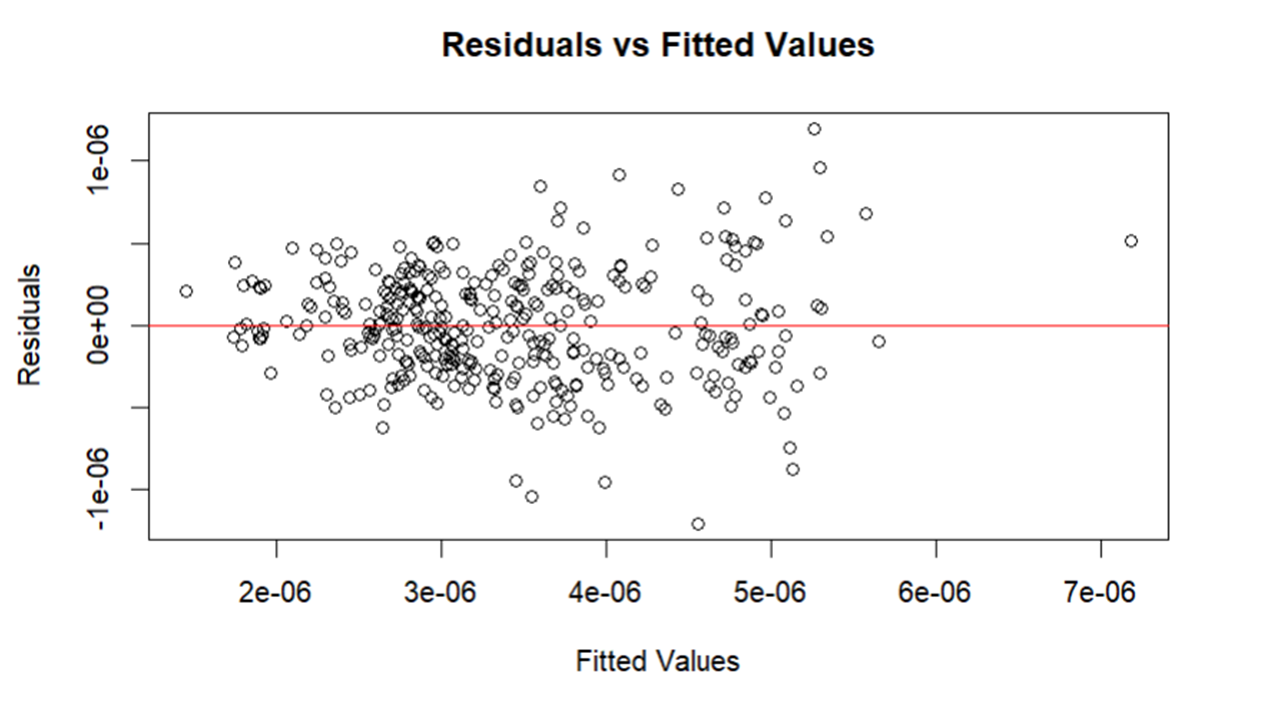
# Appendix E: Figure 5: Residuals vs. Fitted Values for Inverse-Transformed Model



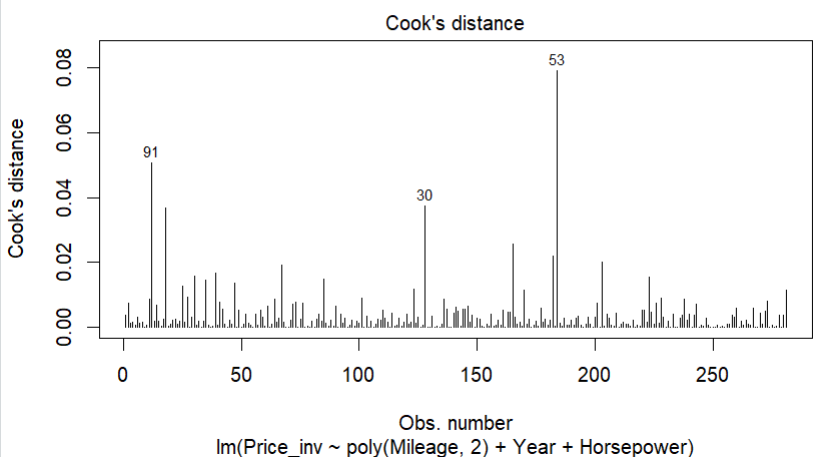
# Appendix F: Figure 6: Normal Q-Q Plot of Residuals for the Inverse-Transformed Model



# Appendix G: Figure 7: Residuals vs Fitted Values for the Inverse-Transformed Model



# Appendix H: Figure 8: Cook's Distance for the Polynomial Inverse-Transformed Model



# References

1.Blocket (https://www.blocket.se/ )

2.Fox, J., & Weisberg, S. (2011). An {R} Companion to Applied Regression (Second Edition). Thousand Oaks CA: Sage. URL: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

3.Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1-26. URL: <http://www.jstatsoft.org/v28/i05/> (caret package)

4.Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. URL: <https://CRAN.R-project.org/doc/Rnews/> (randomForest package)

5.R Core Team. (Year). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

6.Shchutskaya.V.(2021)What is Predictive Performance Models and Why Their Performance Evaluation is Important. InDataLabs. https://indatalabs.com/blog/predictive-models-performance-evaluation-important (2021,May,27)

7.Wikipedia. (2024). Breusch–Pagan test. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Breusch%E2%80%93Pagan\_test. (2024, April 14)

8.Wikipedia. (2024). Generalized additive model. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Generalized\_additive\_model#:~:text=In%20statistics%2C%20a%20generalized%20additive,inference%20about%20these%20smooth%20functions. (2024, April 24)

9.Wikipedia. (2024).Linear Regression. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Linear\_regression. (2024, April 28)

10.Wikipedia. (2024). Polynomial regression. In Wikipedia. Retrieved from <https://en.wikipedia.org/wiki/Polynomial_regression> (2024, Feb 27)

11.Wikipedia. (2024). Random forest. In Wikipedia. Retrieved from <https://en.wikipedia.org/wiki/Random_forest#:~:text=Random%20forests%20or%20random%20decision,class%20selected%20by%20most%20trees>. (2024, Mar 22)

12.Shchutskaya.V.(2021)What is Predictive Performance Models and Why Their Performance Evaluation is Important. InDataLabs. <https://indatalabs.com/blog/predictive-models-performance-evaluation-important> (2021,May,27)

13.Wikipedia. (2023). Statistical Model Validation. In Wikipedia. Retrieved from [https://en.wikipedia.org/wiki/Statistical\_model\_validation (2023](https://en.wikipedia.org/wiki/Statistical_model_validation%20(2023), Dec,11)

14.Wikipedia. (2024).Variance inflation factor. In Wikipedia. Retrieved from <https://en.wikipedia.org/wiki/Variance_inflation_factor> (2024, Jan 23)

15.Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. R News, 2(3), 7-10. URL: <https://CRAN.R-project.org/doc/Rnews/> (imtest package)

16.Zeileis, A. (2004-2020). Econometric Computing with HC and HAC Covariance Matrix Estimators. Journal of Statistical Software, 11(10), 1-17. URL: <https://www.jstatsoft.org/v11/i10/> (sandwich package)