Build a Linear Regression model to predict used car sales for Volvo in Sweden

Testing with blocket.se dataset

Robert Shaw
EC Utbildning
Kunskapskontroll, R Programming
202404

Abstract

This report presents the development of a linear regression model tailored for predicting the resale prices of Volvo cars. Utilizing a comprehensive dataset sourced from Blocket.se and enhanced with real-time vehicle registration data via an API from the Statistics Central Bureau of Sweden (SCB), the model incorporates 753 observations across 10 distinct features, including model year, mileage, fuel type, and engine size. The robustness of the model is highlighted by an R-squared value of 91.5% obtained from validation on a previously unseen test set, demonstrating its high predictive accuracy and generalizability. Additionally, it achieved the lowest Root Mean Square Error (RMSE) among the models tested. These results establish the model as an asset for Volvo dealerships in Sweden, providing a solid basis for informed pricing strategy decisions and efficient inventory management. The integration of real-time data from national vehicle registration databases has proven to significantly boost the efficacy of predictive modeling in the automotive sales sector.

Acknowledgements

I would like to express my sincere gratitude to my teacher, Antonio Prgomet, for his patience and dedication to helping us learn effectively.

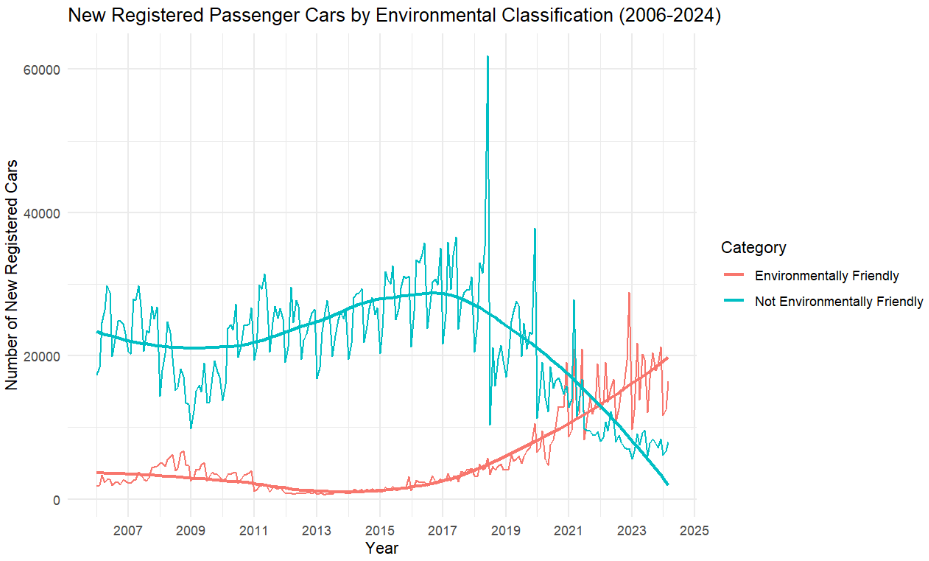
I want to thank all my classmates for helping me in my moments of need to figure out what isn't working and why.

Table of contents

1 Introduction

Regression models are indispensable tools in the quantitative analysis landscape, widely used across various sectors to predict outcomes, identify trends, and make informed decisions. These models are particularly valuable in settings where the relationship between a dependent variable and one or more independent variables needs to be quantified. In business, for instance, regression analyses are pivotal in forecasting sales, understanding customer behavior, and managing resources efficiently. The versatility of these models lies in their adaptability to different types of data and their capability to handle complex, multi-variable relationships, making them fundamental for strategic planning and operational optimizations.

The automotive industry, especially the used car market, exemplifies a sector where regression modeling can provide substantial benefits. Accurately predicting the resale value of vehicles is crucial for dealerships to price their inventory competitively and maximize profit margins. This task, however, is complex due to the myriad factors that can affect a car's market value, such as age, mileage, fuel type, and engine size. To tackle this challenge, this report focuses on developing a multiple linear regression model specifically designed for predicting the sales prices of used Volvo cars. Utilizing comprehensive data from a major Swedish used car sales platform, blocket.se, the model aims to capture the nuanced interplay of various determinants influencing car prices.



In the Swedish Central Statistics Bureau (SCB) graph, we can see here to the right additional insights from national vehicle registration statistics for newly registered cars. Since 2017, while there has been a volatile yet downward trend in fossil fuel car purchases, sales of electric and hybrid cars have been skyrocketing by close to 400%.

This approach not only promises to enhance the accuracy of price predictions but also offers a strategic tool for Volvo dealerships to optimize their pricing strategies and inventory management. By refining the understanding of factors that significantly impact car valuation, dealers can better align their business practices with market dynamics, thereby improving operational efficiency and customer satisfaction in a highly competitive market.

Purpose and research question

The purpose of this report is to construct and evaluate a multiple linear regression model for predicting the resale prices of used Volvo cars. Utilizing a comprehensive dataset from blocket.se and national vehicle registration statistics from SCB, this analysis aims to model the nuanced interplay of various car features on their market value. The ultimate goal is to provide a robust predictive tool for Volvo dealerships to aid in pricing strategy decisions and inventory management, facilitating data-driven approaches in a competitive market landscape.

To guide this investigative effort, the report will seek to answer two pivotal research questions:

1. Can we develop a multiple linear regression model that predicts the resale prices of used Volvo cars with a high degree of accuracy, targeting an RMSE (Root Mean Squared Error) below 20% of the average selling price?
2. Can we develop a multiple linear regression model that explains over 90% of the variance in the selling prices of used cars?

By focusing on these questions, the report will not only enhance the theoretical understanding of factors influencing used car prices but also offer practical insights that can be directly applied in the operational strategies of Volvo dealerships.

1.1 Limitations

We had a very small dataset with limited features that contained less than 1,000 observations and the range/variety of features was relatively narrow. Additional features such as CO2 emissions, fuel efficiency, hybrid or electric vehicle indicator, energy consumption, battery warranty, regenerative braking would be useful given the wider market trend towards environmentally friendly cars.

1.2 Summary

This task begins with an introduction and delineation of the problem. It then goes into more detail on the relevant general and specific theories necessary to understand the scope of the task. This is followed by a description of the experiments and methods used. The results are then presented and interpreted. In summary, the task summarizes the results and suggests avenues for future exploration and refinement of the model.

2 Theory

This section delves into the theoretical underpinnings crucial for understanding the application of multiple linear regression in price prediction and feature selection tasks. We explore baseline logic, normality assumptions, model selection, evaluation, and statistical inference necessary for robust analysis.

2.1 Regression Models

We looked at a range of different aspects of multiple linear regression to handle a price prediction and feature selection task.

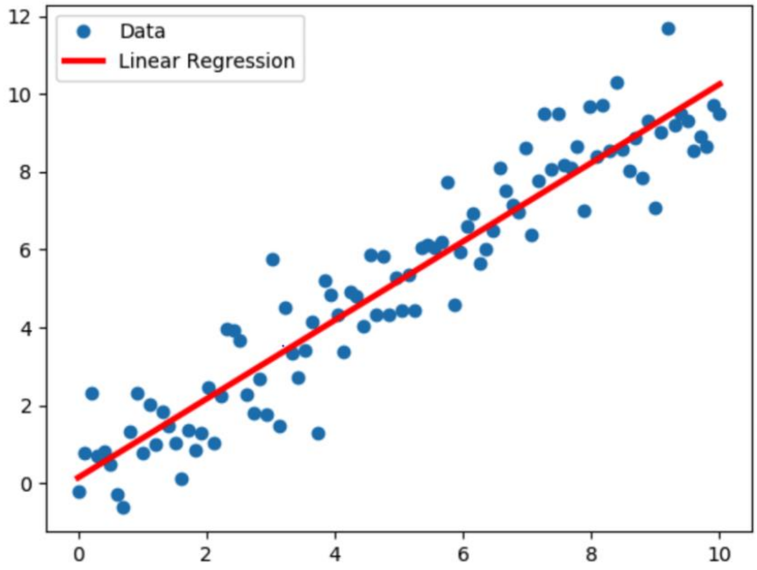
Regression models are a cornerstone of statistical analysis, used extensively to explain relationships between variables and make predictions. These models quantify the impact of one or more independent variables (predictors) on a dependent variable (outcome), by fitting a mathematical equation to observed data. The simplest form, linear regression, assumes a straight-line relationship between predictors and the outcome, but more complex forms like logistic or polynomial regression can model non-linear relationships.

At their core, regression models attempt to draw a line (or a curve, in non-linear cases) that best fits the data points. This "best fit" is usually achieved by minimizing the difference, or error, between the observed values and the values predicted by the model. This methodology not only helps in predicting future values based on the patterns learned from the data but also in understanding the relative influence of each predictor on the outcome.

The power of regression analysis lies in its ability to provide clarity on relationships within data, which can inform decision-making processes across various fields such as economics, medicine, engineering, and more. It is a powerful tool for inference, helping researchers and analysts test hypotheses and derive insights about the forces at play within their data.

2.1.1 Linear Regression models

Linear regression models are fundamental in statistical analysis, allowing us to understand and predict the behavior of one variable based on the behavior of another. In the context of price prediction, linear regression can model the relationship between a car's features (like mileage, year, and engine size) and its selling price. This modeling helps in forecasting prices based on historical data, which is essential for setting competitive prices in dynamic markets (James, Gareth, 2023).



As we can see in the figure here, we are looking for a model that can best fit the relationship between the year it came out on the market and its sales price.

Linear regression models are straightforward to interpret and efficient to use, making them popular for predictive analytics. However, they assume a linear relationship between independent and dependent variables and are sensitive to outliers, which can lead to skewed results if the data is not preprocessed correctly (Montgomery, D.C., 2012).

2.2 Data Preprocessing & Feature Engineering

Missing data can occur due to various reasons such as errors during data collection, entry, or processing. The presence of missing values can bias the results of a statistical analysis if not properly handled, as many algorithms inherently require a complete dataset to perform calculations. Missing data can lead to biased estimates and reduce the statistical power of a study. Removing records with missing values, effective when the proportion of missing data is insignificantly small. Estimating missing values using statistical techniques such as mean, median, mode imputation, regression methods, or more advanced methods like multiple imputation is also useful. Handling missing data is critical to maintain the integrity of statistical inference, ensuring that the conclusions drawn from the analysis are valid and reliable (Roos, Hannah, 2021).

Duplicates in a dataset can arise during data collection or merging processes from different sources. They can distort the analysis, leading to skewed results by over-representing certain information. Duplicates can give undue weight to duplicated observations, affecting the accuracy of predictions and the reliability of the statistical conclusions. Using tools to identify repeated records based on specific criteria and removing them to ensure each data point contributes uniquely to the analysis.

Outliers are data points that deviate significantly from other observations. They could be due to variability in the measurement or experimental errors, and they can indicate experimental errors or novel findings. Outliers can skew and mislead the training process of statistical models resulting in longer training times, less accurate models, and ultimately poorer results. Outliers can be removed, transformed, or left in the data and/or model if appropriate. Proper management of outliers ensures that the model is robust and performs well across the general dataset without being skewed by anomalous data. Residual plots, which visualize the difference between predicted and actual values, can reveal large deviations that might indicate outliers. Studentized residuals, obtained by dividing residuals by their estimated standard errors, are particularly helpful for identifying outliers with unusual combinations of predictor values. Generally, data points with studentized of absolute value greater than 3 are considered potential outliers (James, Gareth, 2023).

2.3 Normality Assumptions

In the context of multiple linear regression, the normality assumption refers to the expectation that the residuals of the model are normally distributed. This assumption is foundational for several reasons. The normality of residuals ensures that the statistical tests for coefficients (like t-tests) are valid. These tests assume that the estimates are normally distributed, which is approximately true if the residuals are normal, particularly when the sample size is large due to the Central Limit Theorem. Also, for Confidence and Prediction intervals to be accurate, the normality assumption is critical. Deviations from normality can lead to intervals that are either too narrow or too wide. The key elements here include the following key aspects.

- **Linearity:** The assumption of linearity is fundamental to linear regression models. It posits that there is a linear relationship between the independent variables (predictors) and the dependent variable (response). This relationship is modeled through a linear equation. Linear models are mathematically straightforward and easier to interpret. The least squares method, which is used to estimate the coefficients, relies on the minimization of the sum of the squared residuals, which can be efficiently computed and has good properties if the linearity assumption holds (James, Gareth, 2023).

- **Independence:** The independence assumption states that the residuals of the model, i.e., the errors, are independent of each other. This is crucial for the reliability of standard errors, test statistics, and confidence intervals. Many statistical tests used in regression analysis, like the F-test for overall significance and the t-tests for individual coefficients, assume that observations are independent to derive their distribution under the null hypothesis. Special care must be taken in time series or spatial data where observations could be correlated across time or space (James, Gareth, 2023).
- **Homoscedasticity**: Homoscedasticity means that the variance of the error terms is constant across all levels of the independent variables. When the variance of the residuals differs at different values of the predictors, the condition is known as heteroscedasticity. Homoscedasticity ensures that the ordinary least squares (OLS) estimator is the Best Linear Unbiased Estimator (BLUE), making it the most efficient estimator among all linear unbiased estimators. If the errors are heteroscedastic, the standard errors for the OLS estimators may be biased, leading to incorrect conclusions in hypothesis testing (James, Gareth, 2023).
- **Normality of Errors:** This assumption states that the residuals of the model are normally distributed. While this assumption is not necessary for the estimation of the coefficients using OLS, it is crucial for conducting hypothesis tests and constructing confidence and prediction intervals. For large samples, the CLT assures that the distribution of the OLS estimators will be approximately normal, irrespective of the distribution of the errors. However, for small sample sizes, actual normality of errors becomes important. The assumption of normality allows for the use of t-tests and F-tests, which rely on the normal distribution (James, Gareth, 2023).

If these assumptions are violated, the results and conclusions drawn from the regression analysis might be misleading or incorrect. Diagnostic tools and remedial measures such as transformations (log, square root), robust regression methods, or even non-parametric approaches might be necessary to address these issues. Understanding and verifying these assumptions through diagnostic plots and tests is crucial in practice to ensure that the linear regression model is appropriate and that the statistical inferences made are valid (James, Gareth, 2023).

2.4 Model Selection Techniques

Model selection techniques are fundamental to improving the predictive performance and interpretability of regression models. These methods help in choosing the most appropriate model among a set of potential candidates by considering different combinations of variables and their complexities. Here's a brief overview of several key model selection techniques:

1. **Best Subset Selection:** This method involves evaluating all possible combinations of predictors to find the model that best fits the data according to a specific criterion, typically the one with the lowest residual sum of squares (RSS) or the highest R2. Best subset selection performs an exhaustive search over all possible models formed by subsets of the predictors. It can be computationally expensive, especially with a large number of predictors (James, Gareth, 2023).
2. **Forward and Backward Selection**: These are stepwise approaches where variables are added or removed one at a time based on their statistical significance in improving the model fit. They are less computationally intensive than best subset selection and provide a straightforward way to understand the impact of adding or removing variables (James, Gareth, 2023).
3. **Interactions:** Including interaction terms in a model allows for the effect of one predictor to depend on the level of another predictor. This is crucial when the relationship between the

predictors and the response is not simply additive. Interaction terms help in capturing the synergy or antagonism between predictors. They can lead to better predictions by modeling the real complexity in data relationships (James, Gareth, 2023).

4. **Regularization Methods**: Ridge and Lasso: These methods add a penalty to the regression model to prevent overfitting, reduce variance, and handle correlation among predictors. Both methods address the bias-variance tradeoff by introducing bias through the penalty to gain a reduction in model variance. They can also help in dealing with multicollinearity among predictors by penalizing the size of coefficients (James, Gareth, 2023).

Each of these model selection techniques offers different advantages and is suited to different scenarios. Best subset selection is ideal when the number of predictors is not prohibitively large, whereas forward and backward selection provide a more practical approach when dealing with a larger number of variables. Regularization methods are particularly useful when dealing with highly correlated predictors or when preventing overfitting is a priority.

2.5 Model Evaluation Metrics

Model evaluation metrics are essential for assessing the performance and suitability of statistical models in capturing the underlying patterns of the data.

- **Adjusted R-squared** is a modification of the R-squared metric that accounts for the number of predictors in the model, providing a measure of how well the model explains the variability of the response data, adjusted for the number of predictors; this prevents the overfitting issues associated with regular R-squared as the number of predictors increases (Brownlee, Jason, 2020).
- **Root Mean Squared Error (RMSE)** quantifies the square root of the average squared differences between the predicted and actual values, offering a clear measure of the model's prediction accuracy where lower values indicate better fit, making it particularly useful for comparing different models or model versions on the same dataset (Brownlee, Jason, 2020).
- **Bayesian Information Criterion (BIC)** is a criterion for model selection among a finite set of models; it is based on the likelihood function and includes a penalty term for the number of parameters in the model, favoring simpler models to avoid overfitting. BIC is particularly useful in the context of model selection, where it balances model complexity and goodness of fit, guiding the selection of a model that is neither too simple to capture the data structure nor too complex to generalize poorly to unseen data (Brownlee, Jason, 2020).

2.6 Statistical Inference

Hypothesis testing in the context of linear regression primarily involves assessing the significance of individual model coefficients to determine if there is enough evidence to conclude that these coefficients are different from zero (or some other value) in the population from which the sample is drawn. When performing regression analysis, each coefficient represents the expected change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables are held constant. Hypothesis testing for these coefficients typically tests the null hypothesis (H0) that the coefficient is equal to zero against the alternative hypothesis (HA) that the coefficient is not zero (James, Gareth, 2023).

Confidence intervals (CIs) provide a range of plausible values for the true population. In the context of regression coefficients, they offer an interval estimate of where the true coefficient is likely to fall with a certain degree of confidence (e.g., 95%).

The confidence interval for a regression coefficient is typically constructed using the point estimate (the estimated coefficient), the standard error of this estimate, and the critical value from the t-distribution that corresponds to the desired level of confidence. If the interval does not include zero, it suggests the effect is statistically significant; the variable contributes to the model in predicting the response variable (James, Gareth, 2023).

Both hypothesis testing and confidence intervals are integral to determining the validity and reliability of the inferences made from a regression model.

3 Method

This chapter describes our step-by-step approach to the project, which includes, data collection and preparation, Exploratory Data Analysis (EDA), model building with cross validation and diagnostic plot testing, until final model selection, testing with unseen data and statistical inference.

3.1 Data collection

The dataset in question comes from the blocket.se and was gathered manually by a group of 5 data scientist students – Keikiet Pham, Melissa Hansson, Jakob Rask, Adrian Andersson Krsmanovic and Robert Shaw - using MS Excel and based on a model goal to collect data from the site that was focused on:

- ✓ Volvo cars in Sweden,
- ✓ Non-leasing cars,
  and
- ✓ based on the fuel type of the car (where the members of the group collected a distinct sample of observations for each type)

We developed a draft template with a random example to be followed and each developed a PoC for our respective fuel type samples. After this, we jointly concluded that our PoCs showed positive signs to be able to create viable multiple linear regression models, we added these:

- ➢ 250 PoC samples (5 members X 50 observations) to another
- ➢ 500 samples (5 members X 100 observations) to give a total of
- ➢ **750 observations for our model**

Keikiet Pham acted as group leader collecting the respective observations into one consolidated Excel file. The group dynamics were good with discussions around varying individual goals leading to a unified decision to target the best-selling brand in Sweden, Volvo and primarily focus on creating a price prediction model for Volvo customers and/or dealerships. If we had more time, we could have targeted collecting more data linked to environmentally friendly cars. My strengths in terms of group work center on listening openly to other perspectives and promoting consensus.

Our group was careful to ensure that we collected as balanced a dataset as possible, for example a similar number of private and dealer listings for each fuel type. By each taking a different fuel type, we also limited the number of potential duplicates in the dataset. The data collection and management process took a considerable amount of time.

The PoC was key using visualizations on a quick-gathered 50-sample set to see that in spite of some serious skewedness in the data, R was able to read the data and the data was sufficiently accurate, consistent, up-to-date, reliable and relevant to the task of producing a basic price predictor regression model for Volvo cars.

The dataset was divided into two distinct sets: a training set that consisted of 80% of the data and a test set that also accounts for 20%. The careful data collection and pre-processing steps ensured that all sets maintain a representative proportion of numerical samples, which contributes to a more robust evaluation of the model.

3.2 Data preprocessing

We first carried out raw **feature selection** to ensure we build a model we can interpret properly.

- We removed brand and first_registration_date straight away as the former was used as part of our model purpose and so does not factor in as a unique value-added feature, and we removed first_registration_date we determined that this is better to replace with model year.

- We also deleted the index as we don't need it anymore beyond data collection.

Trailing spaces: There was a suspected problem w/ horsepower (when converting to numeric (default to NA)), so we decided to check it, and removed all types of whitespace characters and converted to numeric format. We fixed the color feature, so it has just the basic color category and got rid of another problem with trailing spaces in engine size and convert to numeric. We converted the remaining character features to factor (categorical) data types in preparation for our linear regression model.

Missing values: We then moved to handling missing values, fixing the problem with electrical cars that don't have an engine. We solved this by setting 0 for missing values in electric cars, as there were too many observations to delete and it's an important value given it denotes simply that they don't have an engine. We proceeded by imputing median values for other missing fuel types.

We saw that either the seller didn't input values in the advertisement and/or our data team didn't gather the data properly from blocket. Either way, given we only had less than 5% missing values for the remaining columns, we removed the remaining NA rows from the 'vehicle_type', 'drivetrain', 'color', 'transmission' and 'model' columns.

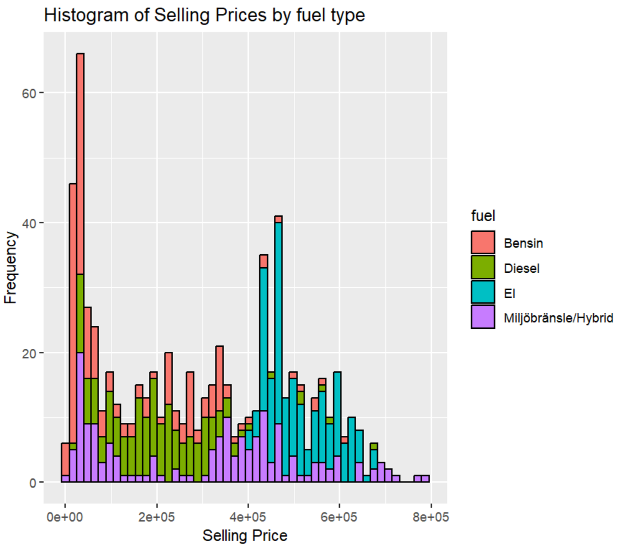
We checked the structure of the dataset and saw that the factor levels had something wrong with the transmission category, where it should have 2 levels not 3. We saw that there seemed to be a break included in 153 observations, that caused R to incorrectly create a new level. We removed them and recreated transmission as a factor with 2 levels.

Duplicates: We checked for duplicates, and we saw that there were 52 duplicates, so we dropped them from the dataset using only unique values.

3.3 Insights from EDA

Exploratory Data Analysis (EDA) began by exploring the relationships between features and checking for anomalies and outliers. We first did a preliminary check for outliers for key numerical and categorical features.

Univariate Analysis: We see a good distribution of selling prices with 75% of these car prices below 425 000 SEK. There is a max value of close to 800 000 SEK and possibly 1 or 2 outliers. The distributions of horsepower, mileage, engine size and model year are all either left/right skewed or non-normal. Fuel and seller types are well balanced given our team designed this in the data collection phase, but there is a certain unbalanced nature to transmission as well as vehicle and model types. Given the extreme non-normal distribution for vehicle types, we decided to create a function to group these together based on 4 simplified and more intuitive and manageable sub-categories. We will deal with the other issues later, as they will both be affected by outliers and will impact our model fit and analysis.



Bivariate Analysis: Here we see quite a few petrol cars have lower sales prices and electrical cars tend to be marketed in higher price range categories. Based on the plots, we can see that both color and model appear to be quite complicated to interpret with many subcategories. We decided to remove color from the dataset and replace model with our newly created and simplified vehicle type column. We see sales prices go down as the number of miles driven increases. There is one particularly extreme outlier, which we assume was a mistake by the seller. Looking closer at the outlier, we see that it's a value over 150,000km. We see from the indexed observation that it's a 2020 car which we deem extremely likely to be an error, so we remove it. A scatter plot shows how selling prices vary based on the model year, with newer cars sold by dealers tending to have higher selling prices. There are a few outliers, but we'll leave these until modeling later. Generally there is an increase in price with an increase in horsepower, but it is not perfectly linear. Engine size doesn't have a clear relationship with selling price, but there is a tendency for the 0 CC or electrical cars to have higher prices and the 2000 CC engine size is quite recurrent and tends to command a higher price range. Again, dealers tend to have higher prices than private sellers. From a correlation matrix and pair plot, we can see that mileage, model year and horsepower seem to have the biggest impact on sales prices, with engine size appearing to trail off in impact.

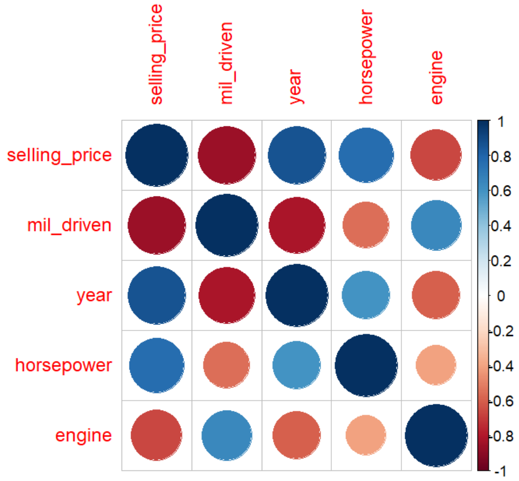
3.4 Train-Test Split

We set aside reserve data approximately 80% of our data for training and 20% as test data for our final best model.

3.5 Training models + Feature Selection/Engineering

Before experimenting with 4 different models using both goodness of fit (R2) and accuracy (RMSE) as our guides, first, we check the reference ranges for RMSE scoring using cross-validation, noting that since the median and mean of our dataset are close and the distribution is somewhat symmetric. We determine that using the mean as a reference is sensible for typical business or economic analyses in the used car sales market where pricing volatility is large; this directly relates error magnitude to an average sales transaction.

Model 1: Our model training process began by using "logic"/theory and was based on our EDA process (see analysis in section 3.3) including our correlation matrix/pair plots, with a view to creating a simple first baseline model, that included mileage, year and horsepower with some transformations. We ran model summary and diagnostic plots.



Correlation Matrix of Numerical Features for Model 1

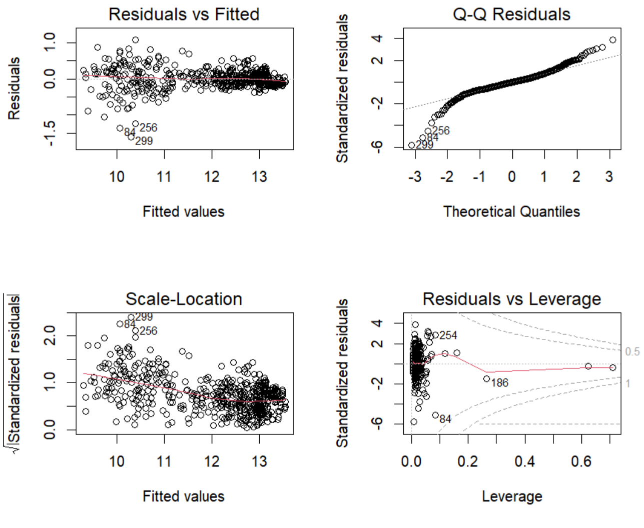
- From this we see that the F-statistic (1426) shows at least one significant value, and moreover that all p-values for the 3 features are statistically significant with p-values < 2e-16 (***), indicating strong evidence against the null hypothesis (i.e. no effect at all). Model year has the greatest impact on selling price, with the coefficient for year being about 8,794, suggesting that each additional year (newer models) increases the selling price by about 8,794, holding other factors constant.
- R-squared: About 88.94% of the variability in selling price is explained by the model. This is a strong fit.
- The diagnostic plots show evidence of non-linearity and mild heteroskedasticity, and possible outliers.

- We see that the Durbin-Watson test is close to 2 so while there are positive correlated residuals, they are close enough to the desired range to leave.
- Multicollinearity isn't a significant problem here with all 3 variables running well under 5. That said, the VIFs between year and miles driven is moderate, which might be due to older cars generally having driven more miles.
- The cross-validation results from Model 1, which showed an RMSE of 67,866 and an R-squared of 0.8905. The high R-squared suggests a good fit in explaining the variance beyond the training data, and the RMSE represents 24.2% of the mean (280,773). This is above our target (under 20%) and may indicate issues such as influential outliers or nonlinear relationships.

Model 2: To enhance model accuracy, we explore more complex modeling techniques and adjustments by adding polynomial to capture non-linear dependencies better, as well as applying a natural log transformation of our target variable to mitigate the impact of skewness and outliers, as well as reducing prediction errors and improving generalization. This balanced approach will ensure not only a good fit but also robust predictive performance across new data. This is precisely what happens with our transformed model 2. It explains a substantially better portion of the variability in the data, with a considerably higher R-squared of 94%, indicating a better fit than model 1. The VIF for polynomial terms of mil_driven and year are relatively high, but the square roots of their generalized VIFs are around 1.5, which, while indicating some multicollinearity, is not extreme. These factors suggest multicollinearity is present but manageable in this context. Horsepower has a VIF close to 1, indicating minimal issues with multicollinearity for this variable.

The diagnostic plots show a major reduction in non-linearity and heteroskedasticity. Even on the cross-validation test, the model shows good generalization with an R2 score of 90%, again an improvement on the first model. The RMSE score of 0.37 is very low due to the log transformation of the target variable. We will look at this again in section 4 below.

We noticed a series of high leverage points and outliers and checked them and removed them based on legitimate irregularities in the data. After removing the high leverage points, the residual clusters have reduced somewhat and the Q-Q residuals plot shows a slightly better straight line of errors.
VIF values for the polynomial terms are relatively high (above 29 for both mil_driven and year), suggesting that there is multicollinearity within these transformations. However, this is common in polynomial regression and we decide that it is acceptable as part of our model's structure. Residuals ranged from a minimum of -1.60426 to a maximum of 1.07357, with the median very close to zero, suggesting that our model generally fits well except for a few outliers. We will create interactions as the basis of our third model to see if this helps. R2 improved to 95% on both the train and CV tests, with RMSE at 0.29 low again, but this is due to the log transformation of y.



Final Model 2

Model 3: Here we run two interactions as the basis of our third model, first on their own with year and then together with year.
The first is Seller Type and Mileage (mil_driven) together with year: The impact of mileage on selling price could vary depending on whether the seller is a private individual or a dealer. For instance, buyers might expect higher mileage cars from dealers to be better maintained than those from private sellers. This interaction is statistically significant and VIFs are low. The interaction term (mil_driven:seller_typePrivat) is positive, suggesting that the negative impact of increased mileage on selling price is less for private sellers compared to other seller types. There is some non-linearity and mild heteroskedasticity in the diagnostic plots, and the residuals range significantly indicating potential outliers or skewness in the distribution of residuals. We decided we will run earlier transformations on this after testing another interaction.

The second interaction term we run is Engine size in cc and Horsepower together with year: This positive interaction term suggests that the negative impact of larger engine sizes on the selling price is mitigated by higher horsepower. In other words, larger engines that also have high horsepower are valued more than large engines with low horsepower. We decide here also that this interaction should be highlighted when predicting prices or analyzing the market for cars with different engine sizes and horsepower ratings. All predictors, including the interaction term, are highly significant, which supports their inclusion in the model. The diagnostic plots show little to no violations of regression assumptions such as non-linearity or heteroscedasticity. The VIFs indicate no multicollinearity. This is a good sign, ensuring that the estimates are reliable.

We decide to build model 3, as an evolved form of model 1 but incorporating both interactions and transforming the target variable with the natural log. This shows a good fit with a 93% R2 on both train and CV tests, low RSE of 0.32 (again due to log transformation of y) and strong statistical significance. VIFs are acceptable. The plots again show no extreme violations of the norms. There are a few outliers but we will leave them in. This model is trying to account for how these factors and their interactions influence the logarithm of the selling price.

- The interaction between mil_driven and seller_typePrivat is significant and negative, indicating the negative impact of higher mileage on selling price is more pronounced for private sellers. The interaction between engine and horsepower is also significant, indicating a complex relationship between these two factors in influencing the selling price.The interaction between engine and horsepower is significant and positive, indicating that the impact of engine size (engine) on the selling price is positively enhanced by the horsepower of the vehicle. This could mean that for cars with larger engines, higher horsepower further increases the selling price, perhaps due to the perceived value or performance benefits associated with more powerful, larger-engine vehicles. This interaction could be particularly relevant in markets where both engine size and horsepower are key selling points, such as with sports cars or luxury vehicles.

Model 4: Our final model was derived using a "best subset selection" method, assessing up to 9 predictors from a total of 13 variables. The analysis revealed a leveling off in the adjusted R2 at about 0.91 after including 5 key predictors: mil_driven, year, horsepower, vehicle type, and fuel. This suggests diminishing returns on model complexity beyond this point. Despite the lack of statistical significance for some variables like Diesel and Hybrid fuels, and Hatchback/Wagon vehicle types, they were retained based on domain knowledge indicating their potential moderating effects, especially in interactions with variables like horsepower. Interestingly, engine size was excluded from the model. This decision was likely due to best subset selection penalizing models for the number of predictors, where the inclusion of engine size did not sufficiently improve performance metrics such as BIC or

adjusted R2 to justify the increased complexity.
The constructed linear regression model with these 5 variables demonstrated a strong fit, explaining 92% of the variability in selling prices on both train and CV tests. The F-statistic further validated the significant impact of these variables on the selling price. The model's VIF values indicated minimal multicollinearity, and diagnostic plots showed no major violations of normality. An RMSE of 59,893, significantly lower than the benchmark model score of 67,866, underscores the model's predictive accuracy.
Electric vehicles were found to command higher resale prices, reflecting a consumer preference for electric mobility driven by factors such as lower running costs and environmental benefits. Conversely, diesel and hybrid vehicles did not show a significant price difference compared to petrol vehicles, highlighting possible market dynamics not captured by the model.

3.6 Swedish Central Statistics Bureau (SCB) API

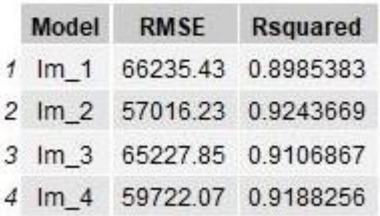
Next, we moved to use the Central Statistics Bureau to get a broader context of the population's demand for cars. We went to their GitHub readme file https://github.com/rOpenGov/pxweb and followed the instructions to access their API. In my R script, I accessed the SCB (Statistics Sweden) API to gather comprehensive statistical data on newly registered cars in Sweden from 2006 through 2024. I utilized the pxweb package, specifically designed for interacting with PX-Web/PC-Axis APIs, which are commonly implemented by statistical agencies, including SCB, for public data dissemination. To specify my request, I constructed a pxweb_query_list. This list detailed several parameters: regional focus, various types of vehicle fuel (like petrol, diesel, etc.), and a monthly breakdown of data spanning over 18 years. The ContentsCode in my query pinpointed the exact statistic related to new car registrations. Upon defining my query, I executed a data retrieval call using pxweb_get, targeting SCB's API endpoint with my specified parameters. The fetched data was then transformed into a DataFrame for easier handling and analysis in R. This setup not only allowed me to systematically access large-scale and long-term statistical data but also facilitated detailed data exploration, including understanding unique vehicle fuel types and the structure of the retrieved dataset. I then created a series of visualizations to better understand the general car sales trends across time as well as the split between environmentally and non-environmentally friendly cars.

4 Results

This section presents a comparison of RMSE scores for our 4 different cross-validated models. By analyzing these results, we assess the performance of the models and their ability to generalize from training to validation and finally to unknown data in the test set. This analysis provides insights into how well each model handles overfitting and its potential for practical application.

4.1 Model Comparison using Cross-Validation

We had log-transformed the target variable - selling_price – for models lm_2 and lm_3. We converted them back to the original scale of the selling price before comparing with lm_1 and lm_4. This was achieved by exponentiating both observed and predicted values, which were originally on a log scale when we trained the model.



	Model	RMSE	Rsquared
1	lm_1	66235.43	0.8985383
2	lm_2	57016.23	0.9243669
3	lm_3	65227.85	0.9106867
4	lm_4	59722.07	0.9188256

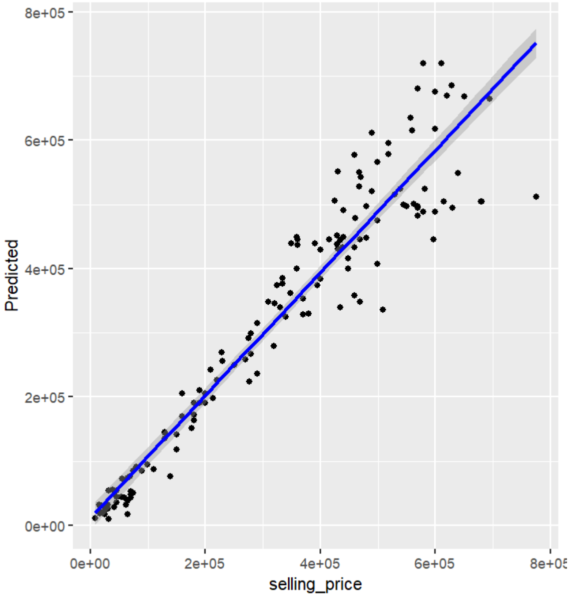
Model Comparison Table for final CV results

4.2 Selection of the Best Model

The best model was selected based on the lowest RMSE obtained from cross-validation results. The RMSE values were compared in a dataframe, and model 2 was found to have the lowest RMSE among all models and is thus determined to be the best model for this dataset. As you can see from the graph, there is a slight deviation – tendency towards heteroskedasticity - from the line as the prices get into a higher range. There is some clustering around the line where the selling price is under 100,000. There are some outliers, particularly again for those cars between 400,000 and 800,000. Overall, there is a narrow confidence band indicating solid model reliability.

4.3 Prediction and Evaluation on Test Data

We made predictions using the best model (lm_2) on our separate test dataset. Since the model was trained on log-transformed data, the predictions are also on the log scale and need to be exponentiated back to the original scale. This step is important for obtaining predictions that are on the same scale as the actual selling prices. The RMSE is then calculated for these adjusted predictions against the actual selling prices in the test data. This RMSE value provides an indication of how well the model performs on unseen data.



4.4 Comparison of RMSE to Mean Selling Price

To put the RMSE into context, it is compared to the mean selling price of cars in the test data. The RMSE on the test data of about 61,059 is approximately 18.83% of the mean selling price, which is about 324,203 SEK. This percentage gives an indication of the error relative to the average value of the target variable. See more on this below in Section 5.

4.5 Hypothesis Testing

Model Significance: The F-statistic and its p-value (< 2.2e-16) demonstrate that the overall model is highly significant. This implies that the explanatory variables collectively provide a significantly better

fit than an intercept-only model.
Coefficient Significance: Significant coefficients, including those for polynomial transformations of mileage and year, as well as horsepower, suggest that these variables have a statistically significant impact on the log of the selling price. For instance, the first polynomial term of mil_driven has a highly significant negative effect, indicating that higher mileage substantially decreases the vehicle's selling price.

4.6 Confidence Intervals (CIs) for Model Coefficients

The narrow confidence intervals for the model's intercept and significant predictors (like the polynomial terms of year and horsepower) highlight the precision of these estimates. For example, the interval for the intercept spans from approximately 11.513 to 11.674, indicating a precise estimation of the baseline log selling price when all predictors are at their mean values.

Range of Influence: The coefficients for year show that newer model years substantially increase the selling price, with the first polynomial term's CI ranging from 16.76 to 19.87, confirming the strong positive impact of newer years on price.

4.7 Confidence and Prediction Intervals

The model's coefficients' confidence intervals (CIs) are quite narrow, indicating precise estimates. For the first car in the subset, the confidence interval (CI) for the predicted selling price ranges from 491,956 to 557,672, reflecting the range where we can be 95% confident the true mean price lies if we were to repeatedly sample under the same conditions. The prediction interval (PI), however, ranges from 301,801 to 909,046 for the same car, highlighting a broader range that accommodates the natural variability and individual prediction errors in selling price predictions. This broader range in PIs suggests there are additional unaccounted factors that could influence the price variations of individual sales.

4.8 Interpretation and Inference

The high R-squared value indicates that the model explains approximately 95% of the variance in log-transformed selling prices, underscoring its effectiveness in capturing the effects of mileage, car age, and horsepower on vehicle pricing. The model provides reliable guidance on how factors like mileage and car year impact selling prices, which can be crucial for stakeholders in pricing used cars or estimating depreciation rates. However, the wide prediction intervals remind us that individual car prices can still vary widely, influenced by factors not included in the model.

5 Conclusions

5.1 Question 1: Can we achieve an RMSE score that is below 20% relative to the average value of the target variable?

We managed to get an RMSE score that is 18.8% of the average selling price of a car on the test data. an RMSE that is less than 20% of the mean is considered good in most business contexts, as it indicates relatively small errors compared to the average transaction value. In the used car market, numerous variables influence pricing, including brand, model, year, mileage, overall condition, market trends, and more. Given this complexity, we can infer the following.

Variability Acceptance: An RMSE percentage, like the 18.8% we have, is very functional, especially if used for initial estimations, broader market analyses, or as part of a tool to help guide rather than dictate pricing decisions.

Model Improvement: If precision is critical, for instance, in a competitive market or for a dealership looking to maximize margins, we could consider refining the model by incorporating additional variables, adjusting the model structure, or using advanced machine learning techniques that can capture more complex relationships and interactions. So, an RMSE of approximately 18% of the mean selling price is very reasonable for a market as diverse and variable as used cars. However, for applications demanding higher accuracy, further model enhancements might be necessary, including using larger datasets and incorporating more fine-tuned features such as CO2 emissions, fuel efficiency, hybrid or electric vehicle indicator, energy consumption, battery warranty, regenerative braking.

5.2 Question 2: Can we develop a multiple linear regression model that explains over 90% of the variance in the selling prices of used cars?

We successfully developed a model with an R2 value of 91.5% on test data, indicating that our model explains approximately 92% of the variance in the selling prices of used cars. This high value suggests that our model is effective at capturing the key dynamics and relationships within the used car market. Importantly, this result was obtained from new, unseen data, underscoring our model’s robustness and its ability to generalize well beyond the specific examples it was trained on. The model’s performance is primarily based on variables such as mileage, year, and horsepower, reflecting a strong and reliable fit between the model and the observed data in real-world scenarios.

However, we must consider the risk of overfitting, particularly because high R2 values can sometimes result from a model too closely fitted to the specific sample data, which might not generalize well to other data sets or future conditions. Additionally, while the model fits our current dataset well, its predictions may vary under different economic conditions or market dynamics not represented in the training set. These considerations underscore the importance of cautious interpretation and application of the model findings to ensure reliability and robustness in varying market conditions.

5.3 Lessons learned and next steps

Insights and Model Experimentation:

Throughout the development of our multiple linear regression model, we focused on understanding the complex dynamics influencing the pricing of used Volvo cars. Our findings affirm that mileage, year of manufacture, and horsepower significantly affect the pricing, with statistical confidence demonstrated by narrow confidence intervals and very low p-values.

During our experimentation, we combined hypothesis testing and predictive modeling to validate the relationships between these variables and to assess the model's prediction accuracy. The dual approach allowed us to refine our model iteratively. We achieved a significant R2 value of 95%, indicating excellent model fit. However, we also aimed to balance this with predictive accuracy, as evidenced by an RMSE score of 18% relative to the average selling price. This balance ensures that while our model explains a high proportion of variance, it also maintains reasonable error margins in predictions. We simplified the model by removing statistically insignificant variables, ensuring that only meaningful predictors were included. This was crucial for maintaining model simplicity and relevance. We implemented cross-validation techniques to verify the model's performance on unseen data, which is critical to ensure its generalizability and robustness.

Addressing Data and Methodological Challenges
To further enhance the model's accuracy and address potential overfitting, we considered advanced techniques such as ridge and lasso regression. These methods are particularly effective in managing multicollinearity among predictors and reducing the model's complexity by penalizing the size of the coefficients. We did not have time to run these techniques in our model.

Incorporating Market Trends and Additional Data
Market trends indicate a significant shift towards electric vehicles, affecting traditional factors like horsepower that might not be as relevant for electric cars. Our findings suggest incorporating new features such as CO2 emissions, fuel efficiency, and electric vehicle indicators to better align the model with current and future market dynamics.

Data Enrichment: Acquiring larger datasets with more nuanced features, including battery warranty and regenerative braking, will allow for a more comprehensive analysis and finer model tuning.
Model Adjustments: Adapting the model to include variables that capture the essence of electric vehicles, and their market value will be crucial. This involves not only adding new variables but also potentially transforming existing ones to better capture the evolving trends in car valuation.

Going forward, we would focus on enhancing the model's predictive accuracy and ensuring it remains adaptive to market changes. This holistic approach will solidify the model's utility in strategic decision-making in the used car market, particularly as we navigate the complexities introduced by the increasing prevalence of electric vehicles.

6 Theoretical questions

Besvara följande teoretiska 7 frågor:

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 Quantile-Quantile (QQ) plot is a graphical tool used to compare the quantiles of two distributions to see if they follow the same distribution.

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?**
   I’d say something like that in machine learning, the focus is mainly on making predictions—forecasting future data points. In statistical regression analysis, you can do both predictions and statistical inference, which means understanding the relationships between variables and making conclusions about the data. For example, regression can tell you the effect of changing one variable on another, so in car data how mileage and model year affect sales, as well as predict future values about sales pricing.

3. **Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?**
   The difference between a "confidence interval" and a "prediction interval" for predicted values is that a confidence interval provides a range within which we expect the true mean of the population to fall, while a prediction interval provides a range within which we expect future observations to fall, i.e. by inputting new data values.

4. **Den multipla linjära regressionsmodellen kan skrivas som: $Y=\beta_0+\beta_1x_1+\beta_1x_2+\ldots+\beta_px_p+\varepsilon$. Hur tolkas beta parametrarna?**
   β represents the expected change in the dependent variable (Y) for a one-unit change in the corresponding independent variable (x), assuming all other variables are held constant. It is like a weighted value that the model estimates based on the importance of each predictor variable.

5. **Din kollega Hassan frågar dig följande: *”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test setom man nyttjar* mått såsom BIC?Vad är logiken bakom detta?”Vad svarar du Hassan?**
   Yes it’s true. BIC helps select the best model by penalizing complexity to avoid overfitting, thus providing a balance between model fit and complexity without needing separate validation or testing.

6. **Förklara algoritmen nedan för ”Best subset selection”**
   Best subset selection" as shown below is an algorithm used in statistical analysis to select a subset of predictors that result in the best performing model. It involves fitting all possible models with all possible combinations of predictors and choosing the model with the best balance of fit and simplicity, often using criteria like BIC or adjusted R-squared.

7. **Ett citat från statistikern George Box är: *“All models are wrong, some are useful.”*Förklara vad som menas med det citatet**.
   Models can’t perfectly represent the real world due to simplifications and assumptions, but still many models can provide useful insights and accurate predictions within certain contexts.

7 Self-evaluation

1. **What challenges you have had during the course of the work and how you have dealt with them?**
   Dealing with the unknown, never-ending coding and constantly feeling like I don't know what I'm doing. Dealing with the messiness of raw data. Talking to/harassing quality classmates like Missy, Jakob and Keikiet helped A LOT!

2. **What grade you think you should have and why?**
   Close to VG given I built a solid model and after a lot of iteration and editing produced a quite insightful report. I also learned a lot by working back-and-forth with my classmates from our data collection group.

3. **Anything you want to highlight for Antonio?**
   I think it was a great exercise. I enjoyed the pain a bit more and appreciated the value of learning how to use raw data to build a not-perfect useful model.

References

1. Brownlee, Jason. (2020). “Probabilistic Model Selection with AIC, BIC, and MDL”. Retrieved April 17, 2024, from: https://machinelearningmastery.com/probabilistic-model-selection-measures/

2. Grover, J & Rishabh, M. (2021). Sculpting data for ML. The first act of machine learning. (First edition). Grover J & Rishabh M.

3. Gunn, Mathew. (2017). " When should one include a variable in a regression despite it not being statistically significant?". Retrieved April 24, 2024, from: https://stats.stackexchange.com/questions/271434/when-should-one-include-a-variable-in-a-regression-despite-it-not-being-statisti

4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.

5. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012) Introduction to Linear Regression Analysis. Vol. 821, John Wiley & Sons, Hoboken.

6. Roos, Hannah. (2021). "Smart Handling of missing data in R". Retrieved April 16, 2024, from: https://towardsdatascience.com/smart-handling-of-missing-data-in-r-6425f8a559f2