Linear Regression Analysis and Modelling

on pre-owned V70 cars

Keikiet Pham

EC Utbildning

Assignment 2

202404

Abstract

This report describes the process and development of a linear regression model to predict advertisement sale prices specifically for Volvo V70 cars on Blocket.se, a Swedish market platform for pre-owned things. The report is focused on identifying crucial predictors and exploring pricing differentials between private and company sales, and includes processes like data preprocessing, feature selection, and model evaluation. In this project, three models are examined. In the final and refined model there are signs of a notably difference in selling prices between private sales and company sales. With the final model it can be statistically proven that while all other predictor variables remain constant, the difference of company seller type makes a difference on average of roughly 16.000 SEK more expensive.

Innehållsförteckning

1 Introduction

According to Statistics Sweden (SCB), responsible for official statistics, Sweden has witnessed a population increase alongside a declining trend in new car registrations. This trend suggests a thriving pre-owned car market. Among Swedish car enthusiasts, the Volvo V70 holds a revered status, remaining the country's most popular car for many years. Renowned for its safety features, reliability, and spaciousness, the V70 has been a preferred choice for families and individuals alike.

Despite Volvo's announcement of the model's discontinuation in 2016, with over 1.25 million V70s sold since its debut in 1996, Teknikens Värld's report on February 3, 2017, highlighted the continued dominance of the Volvo V70 in the Swedish market, with the model boasting the quickest sale times among cars listed on Blocket.se.

This report aims to develop a Volvo V70 prediction model using linear regression to predict advertisement sale prices on Blocket.se. Furthermore, we will explore whether there are differences in prices between private sales and company sales. Through this analysis, we seek to provide insights into the factors influencing sale prices on Blocket.se and how they vary based on the seller type.

1.1 Aim and research questions

This report aims to design a predictive model for Volvo V70 advertisement sale prices on Blocket.se using linear regression. Additionally, it aims to investigate potential price discrepancies between private and company sales.

Research Questions:

- How accurately can a linear regression model predict advertisement sale prices for Volvo V70 cars on Blocket.se?
- Are there notable variations in predicted prices between private sales and company sales?

Measurement of Performance:

The performance of the linear regression model will be measured using metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) value. These metrics will help assess and understand the model's accuracy in predicting sale prices on Blocket.se. A good performance for the model would be reflected in low MAE and RMSE scores, along with a high R^2 value, signaling a strong correlation between the model's predictions and the observed sale prices.

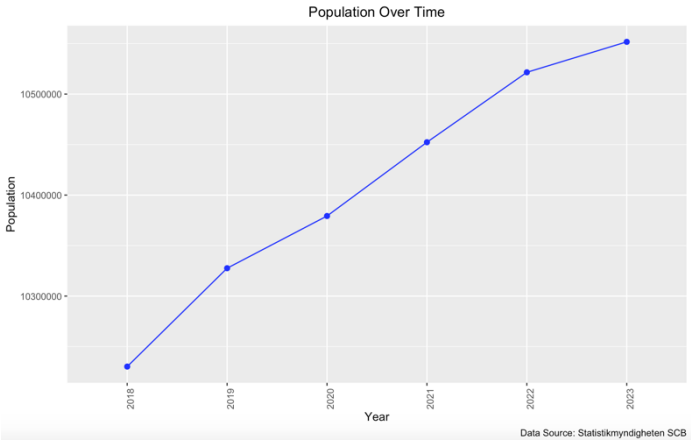


Figure 1 Population growth. From 2018 population has increased by more than 300 000

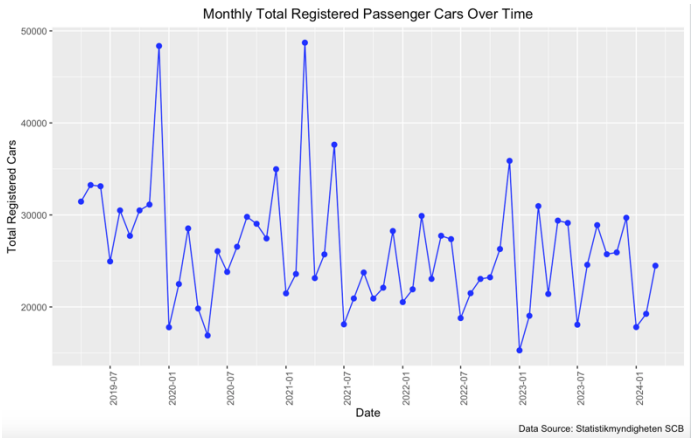


Figure 2 Monthly new car registration. A declining trend in registrations has been observed since 2018."

2 Teori

2.1 7 Potential problem in Regression modelling

Standard linear regression modelling usually provides interpretable results and works on many real-world cases, however several highly restrictive assumptions need to be considered to have valid conclusive interpretable results. While it is important to understand that modelling is a simplified to reality and that the assumptions can practically never be fulfilled it is still important to investigate and deal with the problems.

2.1.1 Linearity Assumption

Linear regression assumes that the relationship between the predictors and the response variable is linear. If the true relationship is far from linear, the model is likely to produce biased estimates and inaccurate predictions. Meaning all conclusions drawn from the model fit may be misleading.

Residual plots are typically useful tools for identifying non-linearity. Ideally the residual plot should not show any clear and distinct pattern which would indicate a problem with some aspect of the model. Techniques like polynomial regression or transformation of predictors such as log(X), $\sqrt{X}$, and X^2 can sometimes address non-linearity (T. Hastie, R. Tibshirani, 2023, p.91-93).

2.1.2 Independance of Residuals

The residuals should be independent of each other. Correlation in the residuals can indicate that there's some pattern left unexplained by the model, suggesting that the model might be missing some explanatory variables. Dependance of residuals typically occurs when there is a time series involved in the independent variables, therefore techniques like time series analysis or including lagged variables can help address correlation in the residuals (T. Hastie, R. Tibshirani, 2023, p.94-95).

2.1.3 Residual Homoscedasticity

This assumption implies that the variance of the residuals is constant across all levels of the predictors. Heteroscedasticity (unequal variance) can lead to inefficient estimates and incorrect standard errors. Diagnostic plots like residual plots or formal tests like the Breusch-Pagan test (bp-test) can help detect heteroscedasticity among residuals (T. Hastie, R. Tibshirani, 2023, p.95-96).

2.1.4 Normality of Residuals

Linear regression assumes that the residuals are normally distributed. Cases of non-normal distribution can affect the confidence intervals and hypothesis tests associated with the model coefficients. Diagnostic plots and Shapiro-Wilk test are tools to help assess Normality of Residuals (T. Hastie, R. Tibshirani, 2023, p.94-95).

2.1.5 Outliers

Outliers are data points that deviate significantly from the rest of the data. Even if outliers may not always have a huge impact on model fit it can still largely impact R^2, RSE and therefore confidence intervals and p-values. Analyzing studentized residuals can help detect outliers, however if suspicion of error in data collection it can be simply removed from the observation. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers (T. Hastie, R. Tibshirani, 2023, p.97-98).

2.1.6 influential Points / High Leverage Points

Influential points, also known as high leverage points, are data points that can have significant impact on the estimated parameters in a regression model and can heavily influence the slope and intercept of the regression line leading to biased parameters estimates and thus affecting the overall

performance of the model. Methods such as Cook's distance measures can help identify influential points (T. Hastie, R. Tibshirani, 2023, p.98-99).

2.1.7 Collinearity & Multicollinearity

Collinearity or multicollinearity occurs when predictor variables are highly correlated with one another resulting in inaccurate predictions and complicated model interpretation. Correlation analysis like Variance Inflation Factor (VIF) and correlation matrices can help identity multicollinearity. VIF values that exceeds 5 or 10 may typically indicate a problematic amount of collinearity. (T. Hastie, R. Tibshirani, 2023, p.100-101).

2.2 Statistical testing

Statistical tests like Shapiro-Wilk and Breusch-Pagan and others can be crucial in data analysis and regression modelling and can play fundamental roles in ensuring validity, reliability and interpretability of regression models.

2.2.1 Rainbow test

Rainbow test is used to check linearity of the relationship between the dependent and variable and independent variable in a regression model where the null-hypothesis is that the relationship between the variables is linear. If the p-value associated with the Rainbow test is less than a chosen significance level, we reject the null-hypothesis, indicating evidence of nonlinear relationships (Vexpower.com Jan, 2022).

2.2.2 Shapiro-Wilk normality test

Shapiro-Wilk test is a test of normality, it tests the null hypothesis that a sample comes from a normally distributed population. Thus, if the p value is less than chosen significance level, we reject the null hypothesis (Wikipedia n.d).

2.2.3 Studentized Breusch-Pagan (bp-test)

The Breusch-Pagan test is used to determine whether heteroscedasticity is present in a regression model. The null hypothesis is that homoscedasticity is present (the residuals are distributed with equal variance (statology.org, Z. Bobbitt, Dec, 2020).

2.2.4 DurbinWatson Test

The Durbin Watson (DW) statistic is a test for autocorrelation among the residuals in a regression analysis. The Durbin-Watson statistic will always have a value ranging between 0 and 4. Where a value of 2.0 indicates there is no autocorrelation detected in the sample. Values from 0 to less than 2 point to positivite autocorrelation and vales from 2 to means negative autocorrelation (Investopedia.com, May, 2023).

3 Data

The data utilized for this project was mainly gathered as a collective effort and consists of 808 data points extracted from advertisements on blocket.se. The data set consists of various details about Volvo cars, such as selling price, seller type, fuel type, transmission type, mileage, model year, car type, driving type, horsepower, color, engine size, registration date, car brand, and model. However, this report focuses solely on Volvo V70 data, ranging car model years from 1998 to 2016.

3.1 Data Description

The data set lists different Volvo V70 cars for sale with their prices, sellers, fuel type, transmission, mileage, model year, body type, and other details.

- Selling Prices ranges from 2000 to 229.900 SEK averaging at 69.900.
- They are sold by both private individuals and companies.
- The cars use either gasoline, diesel, or hybrid fuel.
- Some have manual transmission while others have automatic transmission.
- Mileage ranges from around 15.000 to 61.900 kilometers.
- The cars are mostly from the early 2000s to 2015.
- They have different horsepower, colors, engine sizes, and dates when they were first registered.
- There are 12 rows of duplicates.
- Total of 7 missing values, 4 in engine size, 2 in car type and 1 in transmission type.

3.2 Data Cleaning & Feature Engineering

From the initial 808 data points, only Volvo V70 car models were selected. After removing duplicates, only 152 observations remain. Missing values in car type were imputed (all V70's are Kombi) and one data point of transmission was removed. Although engine size was not used in our models, the data points containing these attributes were retained without modification.

In the data pre-processing phase, categorical features were transformed into factors. The following transformations were applied to the dataset:

1. Säljare (Seller): Converted to a factor variable.
2. Bränsle (Fuel Type): Converted to a factor variable.
3. Växellåda (Transmission Type): Converted to a factor variable.
4. Drivning (Driving Type): Converted to a factor variable.
5. Hästkrafter (Horsepower): Converted to numeric data type.
6. Miltal (Mileage): Converted to numeric data type.
7. Motorstorlek (Engine Size): Converted to numeric data type.
8. Färg (Color): Converted to a factor variable.

These transformations ensure that categorical variables are appropriately encoded for analysis, while missing data are addressed to maintain data integrity and prepare the dataset for further modeling and analysis.

Although potential outliers can be identified, every datapoint is valuable to keep and outliers can potentially play an essential role for future model; therefore, at this stage no points will be excluded.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted in R with the primary objective of uncovering relationships and patterns within the dataset, while also ensuring the data meets the assumptions required for further analysis. Additionally, checks were performed to assess the normality of the response variable, examine the independence of observations, and evaluate linearity between predictors and the response variable.

Below are key take-aways from the EDA:

- Engine size shows relatively strong correlation with mileage, model year, horsepower and selling price. (see figure3)
- Negative polynomial correlation between Mileage and Selling Price, suggesting that as mileage increases, selling price tends to decrease (appendix 1).
- Conversely, a positive polynomial (appendix 2) correlation was identified between model year and selling price, indicating that newer model years are associated with higher selling prices.
- Horsepower showed a weak positive correlation with selling price, implying that higher horsepower may slightly increase the selling price of the car.
- The correlation between car color and selling price was found to be weak to non-existent, indicating that car color may not significantly impact selling price (appendix 3).
- Weak correlation between Fuel type and Selling Price, with potential influence from model year (appendices 5 & 6).
- Analysis also revealed a notable difference in selling price between private sellers and companies, with companies generally pricing higher selling prices (appendix 4).
- Furthermore, there was a substantial disparity in the number of data points between two-wheel drive and four-wheel drive vehicles, suggesting potential implications for analysis and interpretation.

	Miltal	Modellår	Hästkrafter	Motorstorlek	Försäljningspris
Miltal	1.00	-0.36	0.05	0.28	-0.69
Modellår	-0.36	1.00	0.13	-0.61	0.80
Hästkrafter	0.05	0.13	1.00	0.41	0.17
Motorstorlek	0.28	-0.61	0.41	1.00	-0.45
Försäljningspris	-0.69	0.80	0.17	-0.45	1.00

Figure 3. Correlation Matrix. Engine size shows relatively strong correlation with Mileage, Model Year and Horsepower

4 Features

Given their limited data availability, weak correlations with the response variable (Selling Price), and multicollinearity with other independent variables, Driving Type, Car Color, and Engine Size are not prioritized for inclusion. Incorporating these features could potentially introduce noise and unnecessary complexity to the model without significantly improving its predictive power. Our focus remains on selecting the most relevant and impactful features to enhance the interpretability and simplicity of the resulting model.

4.1 Best Subset Selection

In the Best Subset Selection process, the aim is to pinpoint the ideal combination of independent variables that minimize variance in the response variable, Selling Price. We assess various metrics such as Adjusted R^2, Mallows' Cp, Residual Sum of Squares (RSS), and Bayesian Information Criterion (BIC) to evaluate model performance and complexity. Additionally, we apply insights gathered from EDA that there is polynomial correlation between Mileage & Model Year with Selling Price.

To further investigate the potential impact of a 3rd-grade polynomial on our model, polynomial features of grade 3 for selected variables are included. This allows us to capture potential nonlinear relationships more accurately.

4.2 Feature Evaluation

In the evaluation phase, we observe that the inclusion of 3rd grade polynomial variables becomes important in models with 8-11 predictors, indicating a potential concern for overfitting (fitting with many predictors). Therefore, as a proactive measure to maintain model simplicity and mitigate the risk of overfitting, we opt to exclude these variables from our model fit.

In examining the metrics, we observe a trend where RSS decreases as the number of predictors increases. Additionally, Adjusted R^2 peaks with 10 predictors, while Cp is minimized at 9 and BIC at 8, with marginal differences from 7. However, the added inclusion of predictors inherently leads to higher model complexity.

The selection algorithm employed for this evaluation is comprehensive. However, by excluding the 3rd grade polynomial variables, we may find a balance between model complexity and predictive performance, prioritizing the interpretability and generalizability of the model.

	SäljareFöretag	BränsleDiesel	BränsleMiljöbränsle/Hybrid	VäxellådaAutomat	Hästkrafter	poly(Modellår, 3)1	poly(Modellår, 3)2	poly(Modellår, 3)3	poly(Miltal, 3)1	poly(Miltal, 3)2	poly(Miltal, 3)3
1 (1)	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "
2 (1)	" "	" "	" "	" "	" "	"*"	" "	" "	"*"	" "	" "
3 (1)	" "	" "	" "	" "	" "	"*"	"*"	" "	"*"	" "	" "
4 (1)	" "	" "	" "	"*"	" "	"*"	"*"	" "	"*"	" "	" "
5 (1)	" "	" "	" "	"*"	" "	"*"	"*"	" "	"*"	"*"	" "
6 (1)	" "	" "	"*"	" "	"*"	"*"	"*"	" "	"*"	"*"	" "
7 (1)	" "	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "
8 (1)	" "	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"
9 (1)	"*"	" "	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"
10 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"
11 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Figure 4. Best Subset Selection with Model Year & Mileage 3rd grade polynomial.

5 Model Implementation & Fitting

In the Model Selection, Implementation & Fitting phase, we proceed by fitting three multiple linear regression models based on the insights gathered from our previous analysis. For each model, we will examine the model summary, diagnostic plots, and conduct various statistical tests including raintest, Shapiro test, bptest, and Durbin-Watson test.

5.1 Model 1 – insights from best subset selection

In Model1, we incorporate insights from the best subset selection model by including all variables up to the 3rd polynomial as suggested. This includes 7 variables in total:

Predictors: poly(Modellår, 2), poly(Miltal, 2), Bränsle, Hästkrafter, Växellåda

5.2 Model 2 – Simple model with few key predictors

For Model2, we adopt a simpler approach by including only the squared terms of the two main predictors and add Seller type.

Predictors: poly(Modellår, 2), poly(Miltal, 1), Säljare

5.3 Model 3 – Relying on logic and theoretical understanding

In Model3, we rely on logic and theoretical understanding (insights from 3.3 Exploratory Data Analysis) to select predictors. We specifically choose predictors with moderate to strong correlations with selling price while excluding variables that show weaker associations or lack sufficient data for robust analysis.

Car color,and Fuel Type are excluded from consideration due to their relatively weak correlations with selling price and Engine Size due to its multi correlation with other predictors, as indicated by the EDA findings. Additionally Drive is excluded due to lack of observations for four-wheel drive vehicles. We select the remaining crucial features from Best Subset Selection model.

Predictors: poly(Modellår, 2), poly(Miltal, 2), Hästkrafter, Säljare

5.4 Model Comparison

In our model comparison, we investigate seven potential issues across the three models with different variable selections. We compare these models using various metrics to select the most appropriate one for further development and tuning. While some key metrics show similarities and can be fine-tuned across different models, certain factors remain decisive.

Model 1: There are indications of collinearity between Model Year and Fuel Type (appendix 6), where Petroleum fuel type is more prevalent from 1998 to 2005, while Diesel and Hybrid being introduced in 2005. Since Fuel Type is excluded from models 2 and 3, this issue is not present there. Consequently, for this reason Model1 is rejected.

Model 2: Seller Type appears less significant in this model, which is a critical variable for investigation in this project. Therefore, we proceed to explore Model 3 further. Despite Model 3 show signs of heteroscedasticity among residuals, we believe this issue can be addressed with tuning-techniques.

	Model1	Model2	Model3
1) Is there a relationship between the Selling Prices and the predictors?			
P-value of F Statistic	< 2.2e-16	< 2.2e-16	< 2.2e-16
RSE	16790	19809	17522
Adj R2	0,9121	0,8776	0,9042
Std Error	~	~	~
2) What coefficients are not stasticially sifnificant?	Bränsle	Säljare	None
3) Are there any multicollinear features?	Modellår, Bränsle		
4) Is the relationship linear? Check Residual plot			
Visual	~	~	~
RainbowTest	0,7788	0,419	0,52
5) Non-constant variance of error terms			
bptest	0,1566	0,004	0,000704
ncvtest	0,048	0,0000422	0,000048
Visual	~	~	~
6) Correlation of error terms	2,24	2,04	2,2
7) Outlier / High Leverage points	~	~	~

Figure 5. Key metric Model comparison

5.5 Exploring and tuning Model 3 further

Now that we've chosen model 3 to explore further. In this step we address two things:

- Refitting Value range: In the diagnostic analysis our model is predicting poorly in the lower/highest of fitted values (marked in red, figure 6). On a second thought it may be difficult to fit a model that can predict good enough for a as wide range as selling price 2000 to 250.000 without having overly complex model, making inference difficult. Therefore, we limit and retrain our model only with Selling Price data ranging from 20.000 to 160.000.
- Outlier Removal: Additionally, we recognize the presence of outliers, such as datapoint 81, which may impact model performance. In the example of data point 81, the selling price is 200.000 even though Model Year is 2011, we acknowledge it as an outlier and remove it from our model fit.

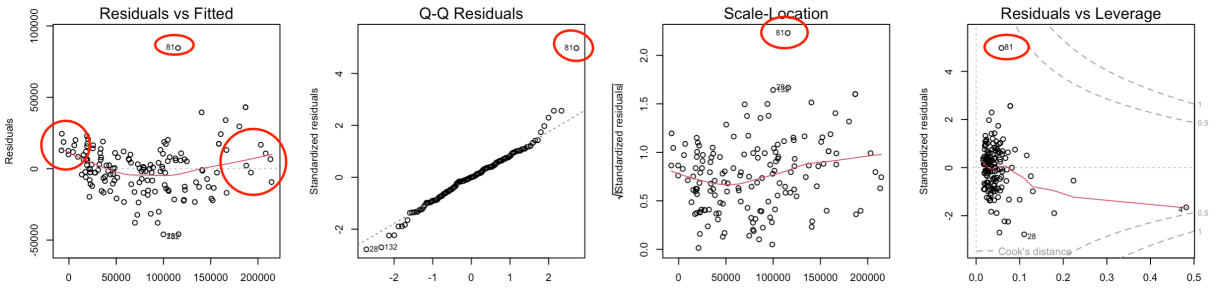


Figure 6. Diagnostic plots of mode3l before tuning.

6 Model Evaluation

After refitting and tuning model 3, we evaluate the performace of our final model. Our model evaluation is done in 3 steps.

6.1 Diagnostic analysis & Statistical tests

In our evaluation of the final tuned Model 3, we double-check our assumptions using various diagnostic plots and statistical tests.

- Residuals vs Fitted Plot: In this plot, we examine the homoscedasticity of residuals. The red line represents the pattern of residuals. Ideally, observations should be evenly scattered above and below the dotted line if the assumption of homoscedasticity holds true. Our analysis indicates a relatively even distribution, supported by a Breusch-Pagan test scoring a p-value of 0.2868, suggesting that statistically the null hypothesis that the residuals are homoscedastic cannot be rejected. Furthermore, no obvious outliers are identified in this plot.

- Normality of Residuals: We investigate the normality assumption of residuals using QQ-Residuals plot and a histogram (figure 7 & 8). Both plots suggest that the residuals follow a normal distribution pattern. Statistical testing with the Shapiro-Wilk normality test confirms this observation, scoring a p-value of 0.799, providing strong evidence in support of our assumption.

- Scale-Location Plot: The flat red line in the Scale-Location plot indicates constant variance of residuals, further validating our model assumptions of homoscedasticity among the residuals.

- Residuals vs Leverage Plot: In this plot, we assess potential high leverage points, which are observations that may have a significant influence on the model fit. While some observations exhibit high leverage, the standardized residuals fall within the range of 0-1, suggesting they may not negatively affect our model significantly. Moreover, the overall leverage is low, indicating minimal impact on the model fit if these observations were to be excluded.

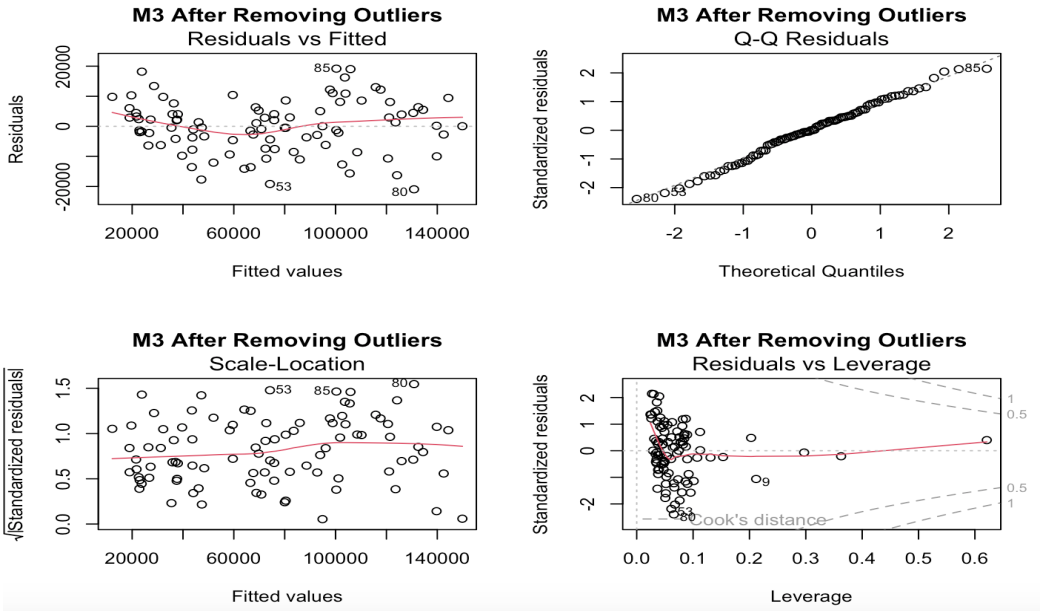


Figure 7. Diagnostic plots on Our Final Regression Model

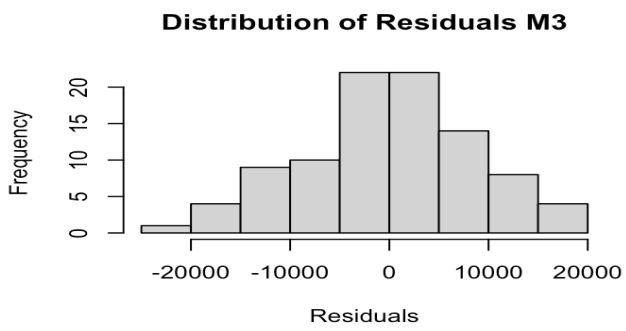


Figure 8. Residuals of M3 tuned looks to follow a Normal-distribution pattern.

To furher strengthen our assumptions, we conduct following statistical tests:

- Raintest: The Raintest statistical test yielded a p-value of 0.1728, indicating that the assumption of linear relationship between the outcome and explanatory variables cannot be rejected.
- DurbinWatsonTest: The Durbin-Watson test resulted in a D-W statistic of 2.009, suggesting independence between the explanatory variables.

Lastly the Variance Inflation Factor (VIF) calculated for the explanatory variables ranged from 1.09 to 1.62, indicating no significant multicollinearity issues among the predictors.

6.2 Predictor Significance

Upon examining the model summary of M3 with now cleaned data, we see that the residuals exhibit a range from -20.965 to 19.178, with a relatively symmetrical distribution around the median, 1Q and 3Q further indicating the residuals follow a normal distributed pattern.

Individually, all explanatory variables show statistically significant t-test results, indicating their importance in predicting the response variable (Selling Price).

The significance of the model as a whole is assessed through the F-statistic. With a low p-value, the null hypothesis that none of the predictor variables are significant is rejected. This suggests that the model, collectively, provides valuable insights in predicting selling price.

```
Call:
lm(formula = Försäljningspris ~ +poly(Modellår, 2) + poly(Miltal,
    2) + Hästkrafter + Säljare, data = car_data_clean)

Residuals:
   Min     1Q Median     3Q    Max
-20965  -5817     28   5421  19178

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          43359.24    5278.20   8.215 1.81e-12 ***
poly(Modellår, 2)1  237062.52   11160.10  21.242  < 2e-16 ***
poly(Modellår, 2)2   82522.38    9476.83   8.708 1.78e-13 ***
poly(Miltal, 2)1   -146804.42    9826.21 -14.940  < 2e-16 ***
poly(Miltal, 2)2     20906.74   10156.21   2.059   0.0425 *
Hästkrafter            132.97      30.84   4.312 4.25e-05 ***
SäljareFöretag       16349.50    2377.84   6.876 8.92e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9065 on 87 degrees of freedom
Multiple R-squared:  0.9484,    Adjusted R-squared:  0.9448
F-statistic: 266.5 on 6 and 87 DF,  p-value: < 2.2e-16
```

Figure 9. Summary of results from Model Fit

6.3 Performance Metrics

Model evaluation involves assessing various metrics to understand how well the models performs to unseen data and make accurate predictions. Since we do not have a validation nor test set, we rely on metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), and adjusted R-squared to evaluate our models.

The Mean Absolute Error (MAE) measures the average absolute difference between predicted selling prices and actual selling prices in our dataset. It provides a straightforward interpretation of the model's predictive accuracy.

In our final model, the MAE is 6914 SEK. This indicates that, on average, the predicted selling prices deviate from the actual selling prices by approximately 6914 SEK.

The Root Mean Square Error (RMSE) is a metric that quantifies the average magnitude of errors between predicted and actual values by taking the square root of the average of the squared differences. Unlike Mean Absolute Error (MAE), RMSE gives more weight to larger errors, making it sensitive to outliers in the data.

In our model, the RMSE value of 8721 SEK represents the square root of the average squared differences between predicted and actual selling prices, taking into account both the magnitude and direction of errors.

Adjusted R-squared is another important metric that assesses the proportion of variance in the dependent variable explained by the independent variables, adjusted for the number of predictors in the model.

In our model, the Adjusted R-squared is 0.9448. This indicates that approximately 94% of the variance in the selling prices can be explained by the independent variables included in the model, after adjusting for the number of predictors.

Lastly, our model shows that the Seller type predictor, significantly influences selling prices. Cars sold by companies have a higher average selling price 16.349 SEK compared to those sold by private sellers given that all other independent variables are constant. With a 95% confidence interval ranging 11.623 – 21.075 SEK and standard error 2377 SEK.

7 Conclussions & Future Improvements

In summary, our final model demonstrates robustness and predictive power, despite lacking validation and test data due to limited observations. We prioritized accuracy by focusing solely on Volvo V70s and price ranges of 20.000 – 160.000 SEK, aiming for a more accurate model tailored to this specific market segment. Instead, we compensate the refined dataset of 94 data points by only having four predictor variables (Model Year, Mileage, Horsepower, and Seller Type) to avoid overfit. The chosen predictors all proved statistically significant. However, despite having a robust model, measuring predicting accuracy on advertising selling prices remains complex, influenced by individual seller valuations, varied selling times, and unmeasured factors like car condition or unique features. The model we have is robust on the current data collected, but future market and selling prices may change due different factors.

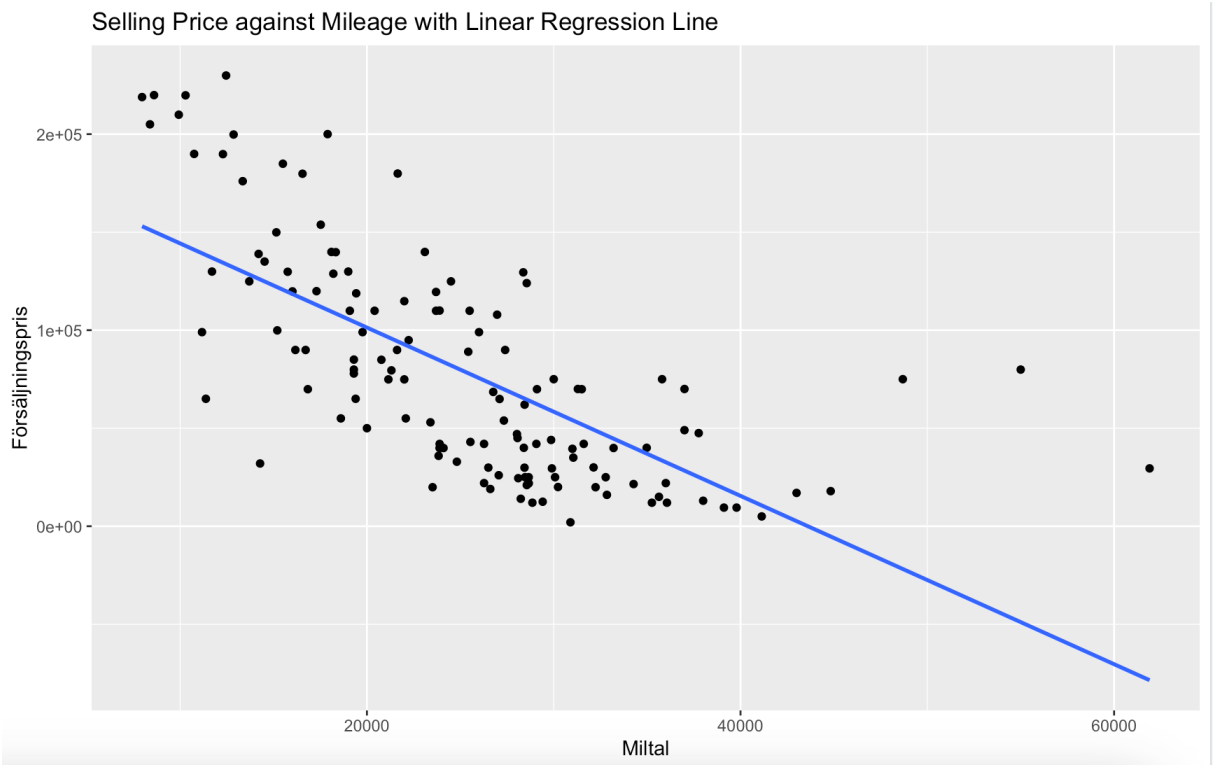
In the end, Model 3, where we applied domain knowledge and insights from exploratory data analysis (EDA), proved to be the better model for refinement. This underscores the significance of leveraging domain expertise and conducting comprehensive EDA to enhance model performance and predictive accuracy.

An interesting finding from our analysis is that there is notable difference in selling prices between private sales and company sales. On way to look at it is that V70 sales ads by companies have an additional cost based solely to the company seller type. This cost is on average 16.349 SEK compared to ads by private sellers. With a 95% confidence interval ranging from 11.623 to 21.075 SEK and standard error of 2377 SEK.

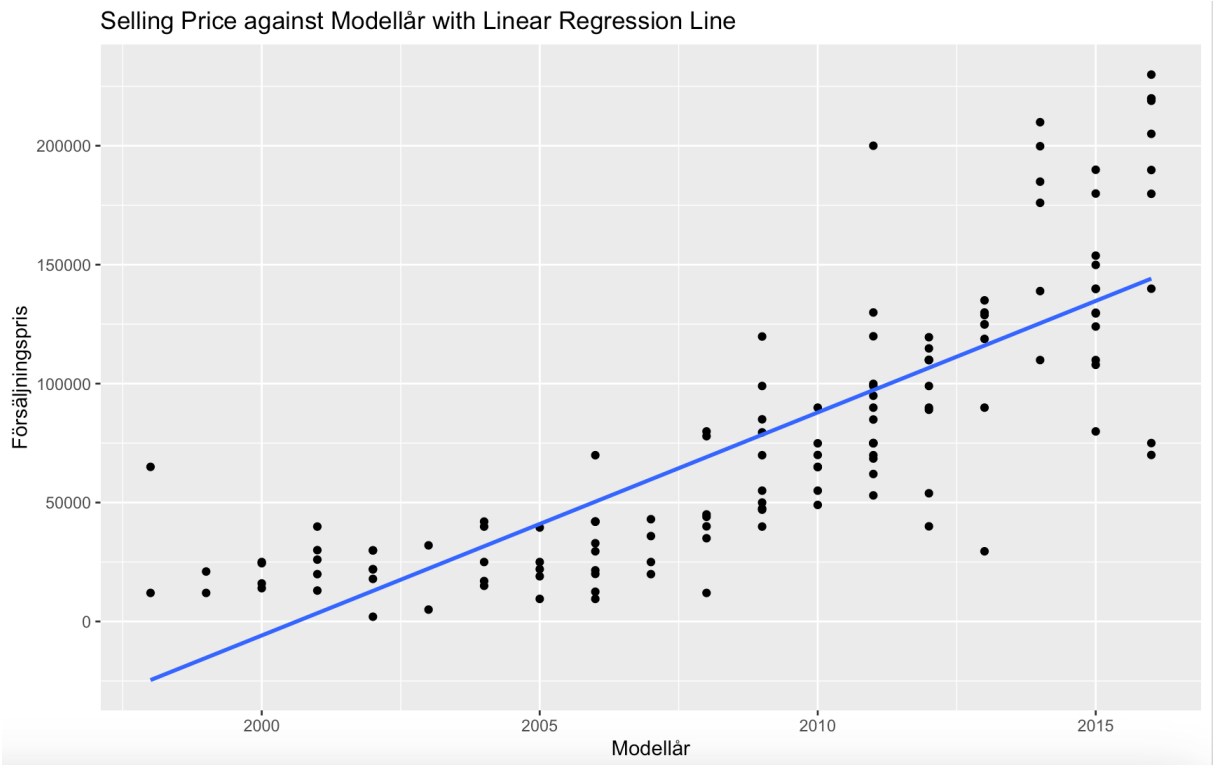
Moving forward, future improvements could involve expanding data collection to include qualitative seller descriptions and exploring additional modeling techniques to further enhance predictive accuracy.

8 Appendix A

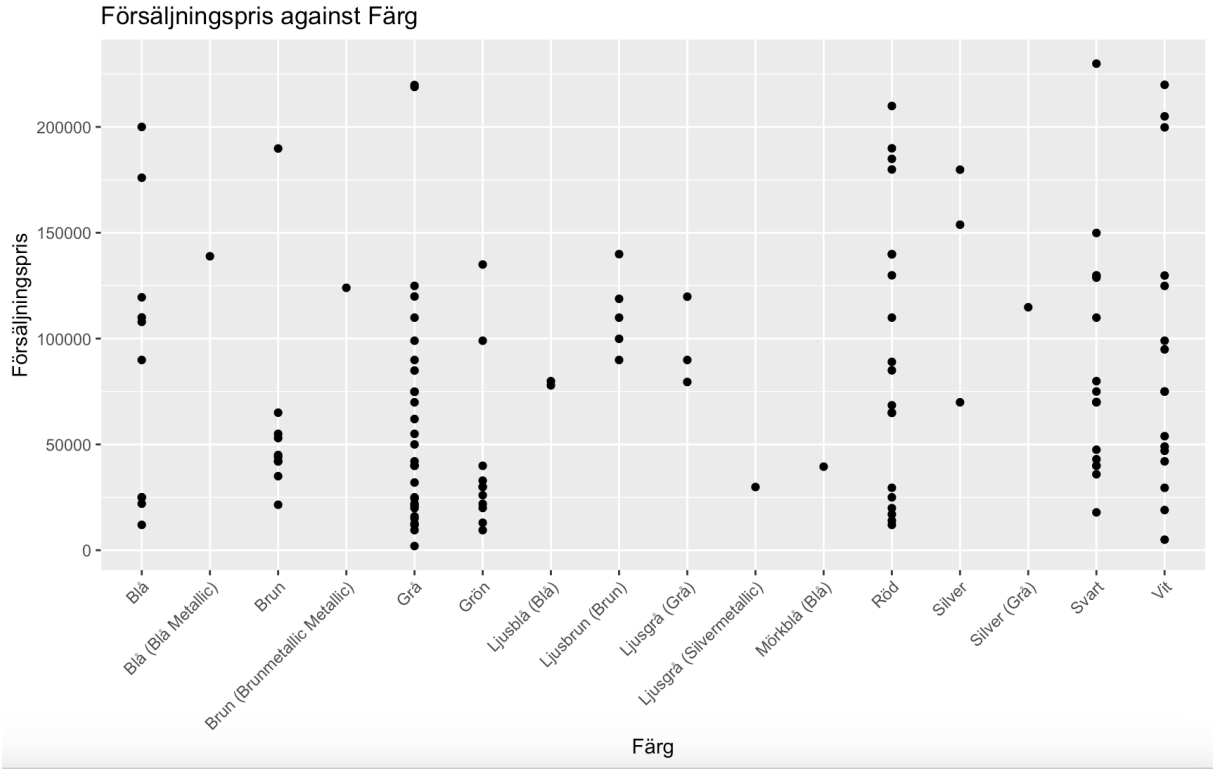
8.1 Appendix 1: EDA: Selling Price against Mileage



8.2 Appendix 2: EDA: Selling Price against Model Year

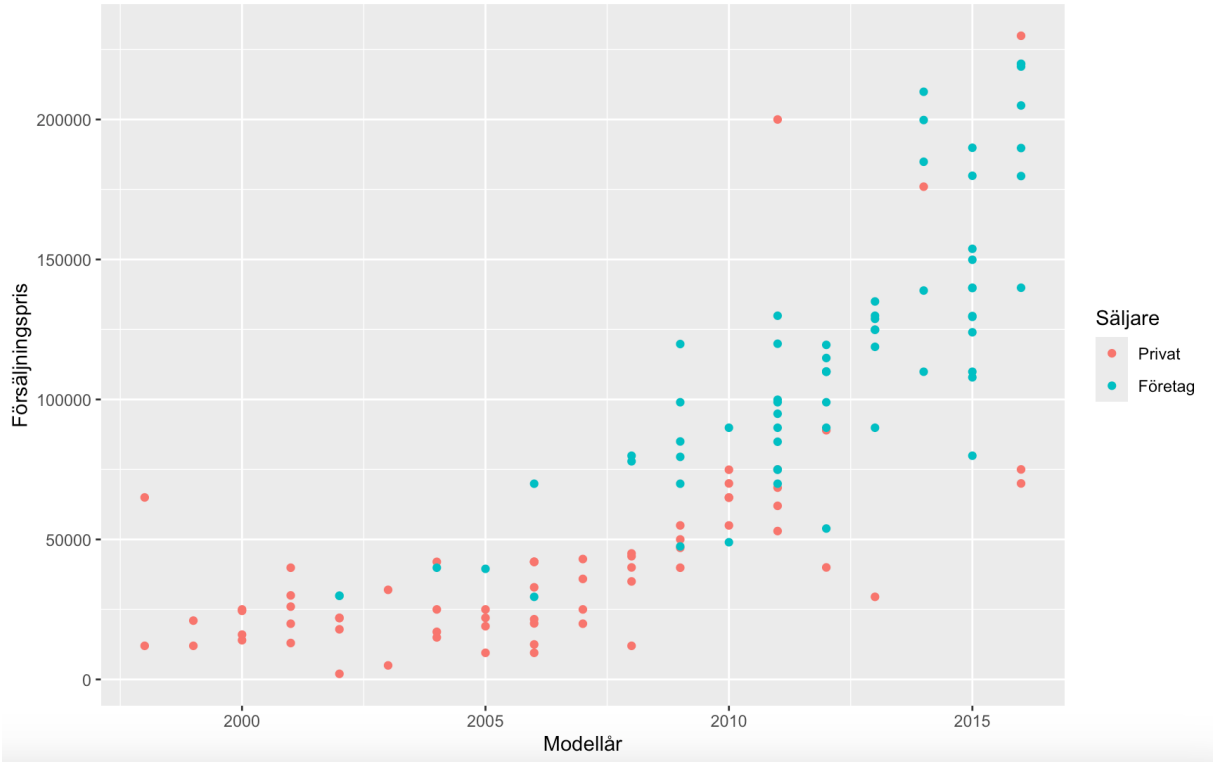


8.3 Appendix 3: EDA: Selling Price against Vehicle Color

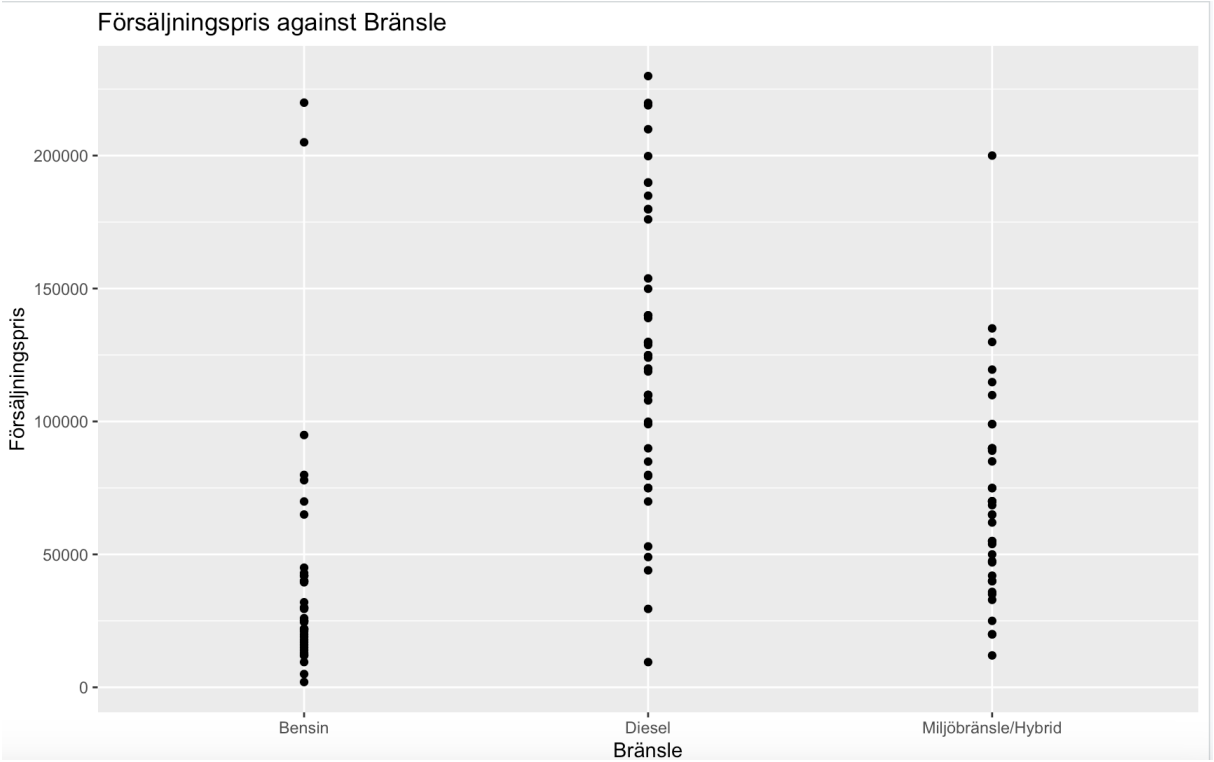


8.4 Appendix 4: EDA: Selling Price against Seller Type

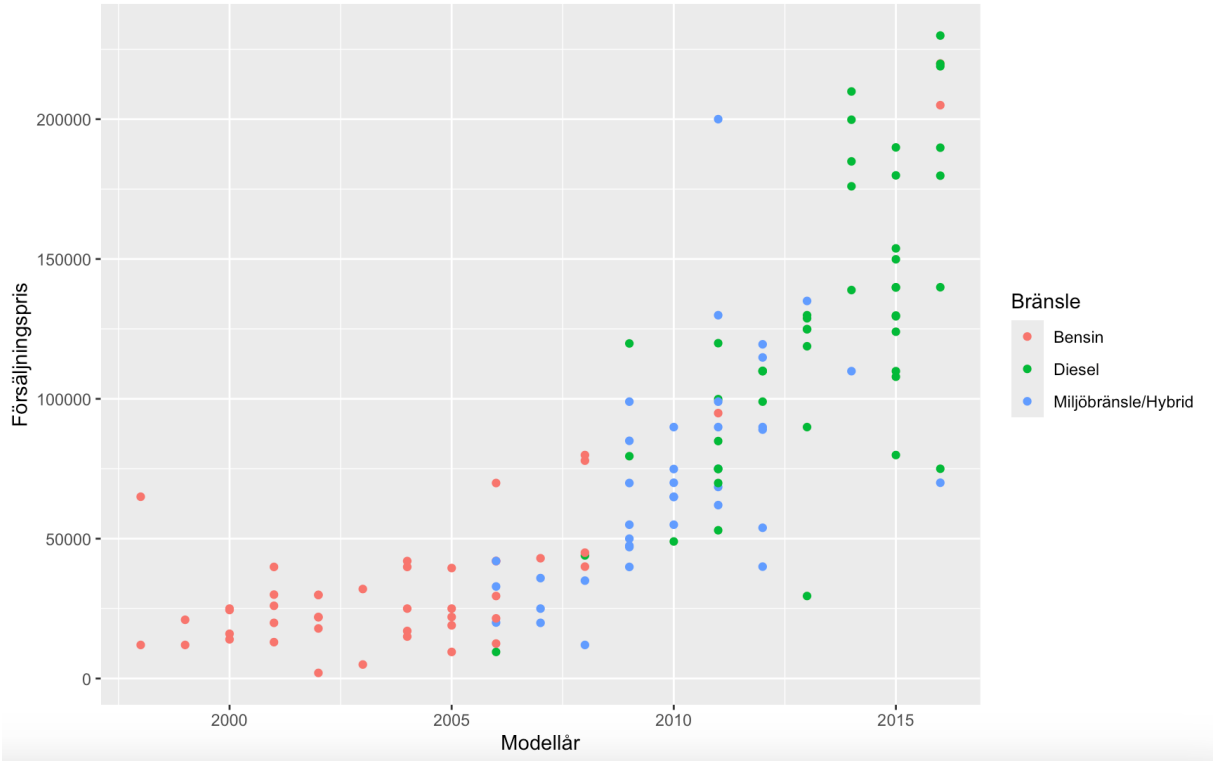
Note: Ads with same model year, Private sellers are usually cheaper than Company sellers.



8.5 Appendix 5: EDA: Selling Price against Fuel Type



8.6 Appendix 6: EDA: Selling Price against Fuel Type



Source list

1) *Gareth, J., Witten, D., Hastie, T., Tibishirani, R. (2023. In: An Introduction to Statistcal Learning – with Applicantions in R) (2nd ed.)*

2) Teknikens värld. (Feb 2017) *Volvo V70 fortfarande populäraste bilen.* https://teknikensvarld.expressen.se/nyheter/bil-och-trafik/volvo-v70-fortfarande-popularaste-bilen-393871/

3) Vexpower.com. (Jan, 2022) Statistical Tests For Linear Regression. https://www.vexpower.com/brief/statistical-tests#:~:text=The%20rainbow%20test%20describes%20whether,will%20be%20greater%20than%20expected.

4) Wikipedia.org (n.d.) Shapiro-Wilk test https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

5) Statology.org (Dec, 2020) *The Breusch-Pagan Test: Definition & Example. https://www.statology.org/breusch-pagan-test/*

6) Investopedia.com (May, 2023) *Durbin Watson Test: What It Is in Statistics, With Examples. https://www.investopedia.com/terms/d/durbin-watson-statistic.asp*