Exploring the use of Linear Regression for Statistical Inference on Used Car Prices



Nike Sandberg Egbalewon

EC Utbildning

Kunskapskontroll – R Programmering

202404

# Abstract

The following report examines the possibility to accurately predict used car prices with a linear regression model and make conclusions on car prices based on statistical inference. Through webscraping, data was collected from the online selling site Blocket.se and three different linear models were trained on the data set. The best performing model was created using Best Subset Selection and a various methods handling issues of heteroscedasticity, multicollinearity and Bias-Variance Tradeoff. The following conclusions were made from the chosen model:

1. A linear regression model with an RMSE score below 50000 SEK could not be achieved during this report.
2. With the use of 15 variables for a linear model, the optimal variables to use were the car brands *BMW, Mercedes, Kia, MG, Nissan, Peugeot, Renault, Toyota* and *Volvo* along with the fuel types *Diesel, Electricity* and *Hybrid, manual gearbox* and *Year* and *Mileage*. Out of the 15 variables, *Year* was the variable with the greatest effect on *Price.*
3. Upon examining the coefficients of the model, following statements were made:
   1. The model indicates that the newer the car, the higher the price.
   2. The car brands BMW, Mercedes and Volvo tend to have a higher price compared to the other car brands represented in the data set.
   3. Electrical and hybrid cars are higher in price according to the model.

Table of Contents

[Abstract ii](#_Toc164972212)

[1 Introduction 1](#_Toc164972213)

[2 Theory 2](#_Toc164972214)

[2.1 Statistical Learning 2](#_Toc164972215)

[2.2 Linear regression models 2](#_Toc164972216)

[2.2.1 Simple Linear regression 2](#_Toc164972217)

[2.2.2 Multiple Linear regression 2](#_Toc164972218)

[2.2.3 Parameter selection 2](#_Toc164972219)

[2.3 Potential Issues to Face when Fitting and Training a Model 3](#_Toc164972220)

[2.3.1 Heteroskedasticity 3](#_Toc164972221)

[2.3.2 Multicollinearity 3](#_Toc164972222)

[2.3.3 Bias-Variance Tradeoff 3](#_Toc164972223)

[2.4 Evaluating Model Performance 3](#_Toc164972224)

[2.4.1 Residual Standard Error and R2 Statistic 3](#_Toc164972225)

[2.4.2 Root Mean Square Error (RMSE) 4](#_Toc164972226)

[2.4.3 BIC 4](#_Toc164972227)

[3 Method 5](#_Toc164972228)

[3.1 Data collection 5](#_Toc164972229)

[4 Result and Discussion 6](#_Toc164972230)

[4.1 Exploratory Data Analysis 6](#_Toc164972231)

[4.2 Estimation of Regression Models 6](#_Toc164972232)

[4.2.1 Model 1: Two Predictors 6](#_Toc164972233)

[4.2.2 Model 2: 29 Predictors 6](#_Toc164972234)

[4.2.3 Model 3: 15 Predictors 8](#_Toc164972235)

[4.3 Model Evaluation 9](#_Toc164972236)

[4.4 Interpretation of Model 3 9](#_Toc164972237)

[5 Conclusion 11](#_Toc164972238)

[References 12](#_Toc164972239)

# Introduction

Despite extensive advocacy against car traffic amongst climate conscious policy makers, activists and non-governmental organisations, cars remain the most used mode of transport in Sweden as of 2022(Sveriges Miljömål). Following this trend, the number of cars in traffic in Sweden have increased by 23 % between 2002 and 2023. In the region of Västra Götaland, the increase during the same time period is 21.5 % (SCB).

Selling and buying cars is a market that shows resilience, and it seems like the used cars market is no exception. According to KVDbil, the leading marketplace for selling used cars in the Nordics, used cars sales has increased by 2,6 % the first quarter of 2024 compared to the same period the previous year. Along with the increase in sales, KVD also reports an increase in prices, especially in diesel fueled cars. However, trends show a decline in prices of electrical cars. (Kvd.se, 2024)

Streamlining the pricing of used cars could be an important tool for sellers, as well as it can provide insight for the buyer as how price fluctuated depending on various parameters. The use of statistical regression analysis can both be used for prediction and therefor an automated way to price cars, but also to provide a deeper understanding for and correlations between parameters that drive pricing.

In this report we will look at both aspects of statistical regression analysis. The purpose of this report is therefor for to provide an understanding of how different variables affect the pricing of used cars in Sweden, as well as determinate whether linear regression models can accurately predict car prices.

Derived from the purpose, the problem statements are as follows:

1. Can a linear regression model predict used car prices with an RMSE less than 50000 SEK?
2. What variables are key variables to explain pricing of used cars, and
3. How do those key variables affect the price of used cars?

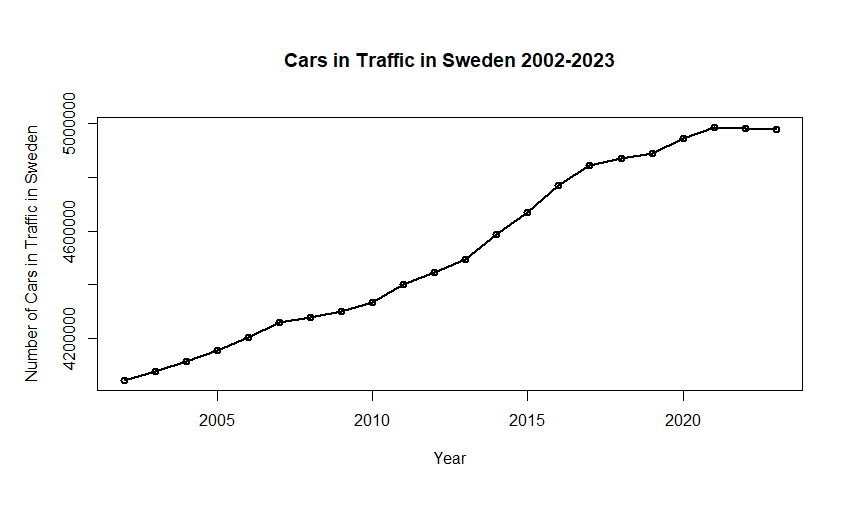


Figure 1. Cars in traffic in Sweden, between 2002 - 2023

# Theory

## Statistical Learning

Statistical learning is a powerful tool to not only observe trends and make predictions, but also create a deeper understanding of data and make conclusions based on that data.

The concept of statistical learning can be traced back to the beginning of the nineteenth century and the development of the “least squares”-method. That was the start of the now widely used form of modelling, linear regression. Since then, and with the development of computing technologies, more complex modelling approaches have emerged making statistical learning an effective method in a variety of fields, from data driven decision making in business to diagnostics in medicine. (James et al., 2023)

## Linear regression models

### Simple Linear regression

Linear regression is a commonly used model in machine learning and statistical analysis. It describes the linear relationship between the dependent variable (y) and one or multiple (multiple linear regression) independent variables (x). in machine learning linear regression is most used to make predictions, whereas in statistical analysis linear regression can also be used to uncover statistical inference. The formula for simple linear regression is:

Where and are the fixed and true coefficients of the model. is the intercept of the model, as it shows the value of Y when X = 0. is the slope of the model and describes the change in Y for each increment step in X. is an error term since a model that perfectly describes reality, in practice, is unachievable. (James et al., 2023)

### Multiple Linear regression

It is possible to use more than one variable for linear regression. This is called a multiple linear regression model. For *p* number of variables, or predictors as they are also called, the multiple linear regression model has the following formula:

Each X signifies the different predictors of the model. The -coefficient is the intercept of the model, just as in the simple linear regression model. describes the average increase or decrease in Y for each unit increases, given that all other predictors are constant. describes the average increase or decrease in Y for each unit increases, given that all other predictors are constant, and so on for each -coefficent up to . As in simple linear regression, is the error term of the model. (James et. al., 2023)

### Parameter selection

#### Best Subset Selection

Best Subset Selection is a method used for choosing parameters for a model. For *p* number of predictors, the method generates a vast number of models used for comparison before selecting, the best one for each number of variables up to *p* predictors. This is because the method calculates the best variables out of all possible combinations - for each possible number variables. The method is for this reason not suitable for data sets with a large number of variables. Not only will the method be time consuming in such an instance, but there is also a possibility of getting a well performing model due to sheer chance. (EducationTopicsExplained, 2024).

## Potential Issues to Face when Fitting and Training a Model

### Heteroskedasticity

One key assumption when estimating the -coefficent, is that the variance of the residuals is constant. Otherwise, the -coefficent estimates, along with confidence and prediction intervals and hypothesis testing, will not be accurate. When the variance of the residuals is non-constant, heteroskedasticity occurs. One possible way of handling heteroskedasticity in a model is through a logarithmic transformation of the dependent variable. (EducationTopicsExplained, 2024).

### Multicollinearity

Multicollinearity in regression models occurs when two or more predictors have a strong correspondence. This correspondence may cause problems for the model since the assumption for estimating the -coefficents of the model is that the predictors are independent from one another.

Calculating the Variance Inflation Factor (VIF) can be used to detect multicollinearity. The VIF score scales between 1 to infinity. The higher the score, the higher multicollinearity is indicated. A score higher than of 5 or 10 is often used as a reference point to determine multicollinearity. (EducationTopicsExplained, 2024)

### Bias-Variance Tradeoff

Bias-Variance Tradeoff refers to the understanding of the prediction errors from a model. While bias can be explained as the error the model makes, between the true and predicted values, high variance is an indication on the model’s complexity. "A model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before.” (Towards Data Science, 2018)

For this reason, it is important to be mindful of the bias-variance tradeoff when evaluating and selecting a model. High complexity and a large number of predictors may not be beneficial. Simpler models are often preferred.

## Evaluating Model Performance

### Residual Standard Error and R2 Statistic

Residual Standard Error (RSE) is a useful metric to evaluate a model’s performance. It estimates the standard deviation of the model’s residuals, i.e. the difference, at any given point, between the predicted value and the true value. Since RSE is a measurement of a model’s error, the aim is to get as low of a score as possible.

R2, also called the coefficient of determination, is a measurement for “the proportion of variance in the dependent variable that can be explained by the independent variable”. Since R2 is calculated on the training data, the inclusion of additional variables to the model will result in a as high or higher score. This means a higher number of variables will be favored by using R2 as a metric, even though simpler models are often preferred. Adjusted R2 is a solution to this problem since it includes a penalty for the number of variables, which adjusts the score and give a more accurate metric when comparing models with a different number of variables. (Corporate Finance Institute)

### Root Mean Square Error (RMSE)

Root mean square error can be a very interpretational way of measuring the model’s performance, especially if the purpose of creating the model is to make predictions. The formula for the metric is:

The formula generates an average measurement of the “error” of the model, i.e. an average difference between the predicted value and the true value. (James et al., 2023)

### BIC

BIC is a metric used to evaluate models, without the need to split the data set into training, validation and test sets. The metric estimates the training error, for which a smaller size of the BIC value is desirable. Compared to other metrics used for the same purpose, such as Cp and AIC, BIC is higher for models with a higher number of predictors. (EducationTopicsExplained, 2023)

# Method

## Data collection

The data used for the report was collected from the Blocket website using webscraping. The webscraping was limited to cars being sold in Västra Götaland, and the full data collection was saved in an Excel file.

An extensive cleaning of the data in Excel was initiated were the information collected was organized in columns named *Brand, Model, Year, Mileage, Fuel, Gearbox* and *Price*. Faulty ad information or rows with cars that was advertised for lease instead of purchasing was removed from the data set.

Luxury brands like Porsche, Bentley and Lamborghini were chosen to not be included in the data set to prevent outliers, that could negatively affect the RMSE

Once the data set was collected and cleaned, it was loaded into R Studio for following methodical steps to be taken:

1. Exploratory data analysis (EDA)
2. Estimation of regression models
3. Problem diagnostics
4. Model Evaluation
5. Interpretation of chosen regression model

During the second step, the data set was divided into a training set, a validation set and a test set, with a 60 %, 20 % and 20 % split. Three approaches of estimating regression models were used. The first approach was a assumed a linear relationship between the predictors “Year” and “Mileage” and the response “Price”. This correlation was fit into a linear model, using the least squares method of calculating the best fit for the data by calculating the coefficients to minimize the square vertical distance from the data points to the linear model.

The second approach was to fit the entire set of predictors in a linear model, followed by the third approach using best subset selection to choose the optimal number of variables for fitting a model.

Following steps were explored to improve the performance of the models:

* Logarithmic transformation of the dependent variable
* Handling of multicollinearity through removal of predictors
* Best Subset Selection

# Result and Discussion

## Exploratory Data Analysis

The loaded data set had a dimension of 5232 rows and 7 columns. Three of the columns contains numerical information and four of the columns contains categorical information under the class “character”. The data set contains no NULL values.

## Estimation of Regression Models

### Model 1: Two Predictors

The R2 score of the Model 1 shows that a relatively low proportion (40 %) of variance in car prices could be explained by the model. Both of the variables in the model were significant, with p-values close to zero. The inclusion of more variables could create a model with a higher R2- score.

However, does both of the variables show significance, which indicates that Year and Mileage are key variables to predict Price, but additional steps are needed to create a more accurate model.

### Model 2: 29 Predictors

Model 2 included all six predictors - *Brand, Model, Year, Mileage, Fuel, Gearbox.* Four of the six predictors contained categorical values, causing dummy variables to be created for each unique value upon fitting the predictors to the model. This increased the number of variables from 6 to over 200.

Model 2 generated a higher R2 as predicted, due to the vast increase in predictors. Looking instead at the Adjusted R2 value, it still showed a higher value than Model 1. However, there seemed to be some issues of multicollinearity.

Analyzing the coefficients of the model, 13 were not defined. This is an error that can occur in the presence of multicollinearity. Looking at the VIF score, *Brand* and *Model* received infinite VIF scores, which implies perfect multicollinearity. Upon this stage, *Model* was chosen to be removed as a predictor from the model. With *Model* removed as a predictor, Model 2 had no predictors with a VIF score higher than the chosen limit of 5. One could also argue for a correlation between *Year* and *Mileage*, since older cars tend to have a travelled further than newer cars. But since both predictors were significant and had a multicollinearity score of under 5, they were kept in the model. *(See Figure 2 and 3)*

Plotting the model, however, shows tendencies of heteroskedasticity and residuals deviating from Normal Distribution. Using more variables creates a more complex model with lesser interpretability, which is often something is undesirable due to the Bias Variance trade off.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figure 2. VIF Score Before Removing Model as a Predictor

En bild som visar text, Teckensnitt, skärmbild, nummer

Automatiskt genererad beskrivning

Figure 3. VIF Score After Removing Model as a Predictor

### Model 3: 15 Predictors

With the reduction to 29 variables, Best Subset Selection was assessed to be the primary approach to decide the optimal number of predictors for Model 3. Plotting the Adjusted R2 score for each number of predictors, the diagram shows that the highest value is generated from including all of the 29 predictors. However, after 15 predictors, there is no vast change in the score. The value only decreases from 0.6311 to 0.6149. (*See Figure 4*)

A further investigation of the diagnostic plots for the model showed tendencies of heteroscedasticity and residuals not being normally distributed. A constant variance of the residuals and normal distribution are assumptions for linear regression model to ideally be fulfilled. A logarithmic transformation of the dependent variable gave an improvement of the result when looking at the diagnostic plots. The Adjusted R2 value increased from 0.6149 to 0.7271 by this transformation.

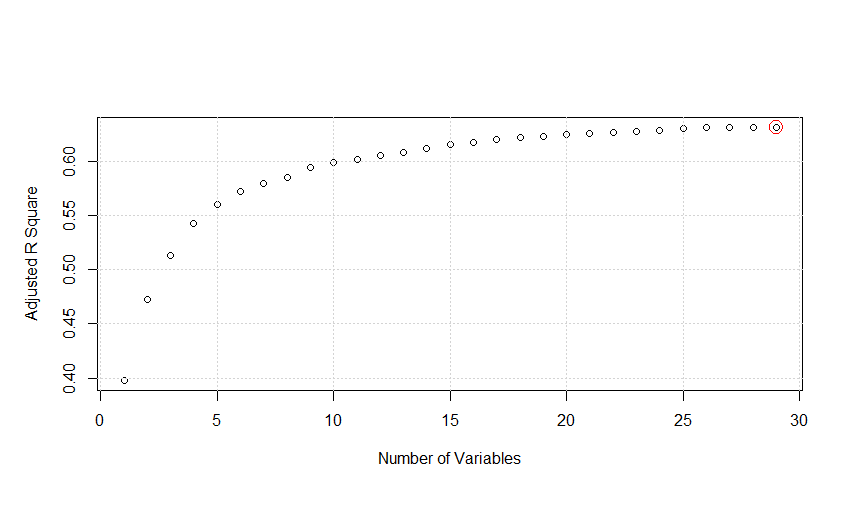


Figure 4. Plot of the Change in Adjusted R2 dependent on the Number of Variables

En bild som visar text, skärmbild

Automatiskt genererad beskrivning

Figure 5. Residuals vs Fitted Values Before Logarithmic Transformation

En bild som visar text, skärmbild, diagram

Automatiskt genererad beskrivning

Figure 6. Figure 6. Residuals vs Fitted Values After Logarithmic Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **RMSE** | **Adjusted R2** | **BIC** |
| Model 1: 2 predictors | 161040.4 | 0.4001977 | 83966.164 |
| Model 2: 29 predictors | 127194.8 | 0.6310980 | 82630.645 |
| Model 3: 15 predictors | 117411.2 | 0.7270532 | 1420.318 |

Figure 7. Performance Metrics for the Models

## Model Evaluation

All three models were evaluated using two approaches. The first approach used the validation set to calculate RMSE. The second approach used Adjusted R2 and BIC on the

training data to estimate test error. All three metrics were in favor of Model 3, which is why it was chosen as the model to use for the purpose of the report. *(See Table 1)*

The result of the RMSE on the test set showed a higher value of 118575.5, which in the context of the report can be interpreted as the average error of the model when predicting prices is 118575 SEK.

When testing the model on new data it performed better, but still making prediction errors of over 50000 SEK on the two of the three cars in the new data set.

## Interpretation of Model 3

Hypothesis testing on Model 3 was performed, with the null hypothesis that the predictors have now effect on the target variable *Price.* The summary of the model shows close to zero values on all predictors, well below the chosen limit of 0.05 %, which allows us to reject the null hypothesis.

Confidence intervals and prediction intervals with the level of 95 % certainty disclosed the lower and upper limit of the predictions for the new data. The true value would be found in all corresponding intervals.

Examining the coefficients, it shows that *Year* is the predictor with highest effect on car prices. BMW, Mercedes and Volvo are the three car brands that according to the model increases the car prices the most out of the brands, if all other conditions are constant.

*Manual* is a dummy variable for the Gearbox variable. The coefficient -3.77 indicates that if all other variables are constant, a car with a manual gearbox cost less than a car with an automatic gearbox. This could, however, also be correlated to the fact that newer cars have an automatic gearbox in greater occurrence than older cars. Hence, could this variable be in interaction with the *Year* variable.

Since the chosen model generated a high RMSE score in relation to the purpose of the model, it is not useful for predictions. However, it could be used to create understanding for customers, either potential sellers or buyers, what variables affect the price of a used cars and give an inclination how they affect the price. The exact figure of the coefficient should not be taken for absolute truth however, due the high RMSE, but looking at whether the coefficient is positive or negative could be useful for understanding how, rather than how much, the different variables affect price.

# Conclusion

1. A linear regression model with an RMSE score below 50000 SEK could not be achieved during this report.
2. With the use of 15 variables for a linear model, the optimal variables to use were the car brands *BMW, Mercedes, Kia, MG, Nissan, Peugeot, Renault, Toyota* and *Volvo* along with the fuel types *Diesel, Electricity* and *Hybrid, manual gearbox* and *Year* and *Mileage*. Out of the 15 variables, *Year* was the variable with the greatest effect on *Price.*
3. Upon examining the coefficients of the model, following statements were made:
   1. The model indicates that the newer the car, the higher the price.
   2. The car brands BMW, Mercedes and Volvo tend to have a higher price compared to the other car brands represented in the data set.
   3. Electrical and hybrid cars are higher in price according to the model.

# References

James, G., Witten, D., Hastie, T., Tibshirani, R (2023). *An Introduction to Statistical Learning,* (2nd edition). Springer

Corporate Finance Institute, *R-Squared,* <https://corporatefinanceinstitute.com/resources/data-science/r-squared/>

Kvdbil, *Begbilsrapport: Fler affärer och högre priser i första kvartalet*, (April 17, 2024), <https://www.kvd.se/artiklar/livet-med-bil/begbilsrapport>

Sveriges Miljömål*, Resvanor uppdelade på färdsätt och kön*, <https://www.sverigesmiljomal.se/miljomalen/frisk-luft/resvanor/>

Toward Data Science, *Understanding the Bias-Variance Tradeoff* (May 21, 2018) <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>

EducationTopicsExplained, *StatsLearning Chapter 6 part 4,* (April 5, 2023), <https://www.youtube.com/watch?v=f10jUj3oqG4&list=PLgzaMbMPEHEydThXS_bCyU5pYTHL7iayF&index=5>

EducationTopicsExplained, *Linjär Regression (*March 19, 2024) <https://www.youtube.com/watch?v=NcxMuCG6FS8&list=PLgzaMbMPEHEyLy3NJ8tZqHBzoVcZlowX4>

API from SCB can be found at: <https://github.com/nikesandberg/DS23.git>

Code used for the report can be found at <https://github.com/nikesandberg/DS23.git>