Regression Modelling and Analysis with Multiple Linear Regression



Erik Fredrixon

EC Utbildning

R Kunskapskontroll

2024–04

# Abstract

In the work for this report two multiple linear regression models were created. The models were trained, validated, and tested. The data collected for the models comes from real passenger vehicles data collected from Blocket.se, using a web scraper. The first model was a default multiple linear regression model, and the second model had the Y variable logged. The second model with the log(Y) variable preformed the best overall, with a RMSE of 39 000kr and Adjusted R^2 of 0.803. The second model also had the better diagnostic plots. This shows that it is important to take many different things into account when evaluating models in Linear Regression projects.

Table of Contents

[1 Introduction 1](#_Toc164954693)

[2 Theory 2](#_Toc164954694)

[2.1 Linear Regression 2](#_Toc164954695)

[2.1.1 Simple Linear Regression 2](#_Toc164954696)

[2.1.2 Multiple Linear Regression 2](#_Toc164954697)

[2.2 Evaluation Metrics 2](#_Toc164954698)

[2.2.1 Residual Sum of Squares (RSS) 2](#_Toc164954699)

[2.2.2 Residual Standard Error (RSE) 3](#_Toc164954700)

[2.2.3 Adjusted R-squared 3](#_Toc164954701)

[2.2.4 Bayesian information criterion (BIC) 3](#_Toc164954702)

[2.2.5 Root Mean Square Error (RMSE) 3](#_Toc164954703)

[3 Methodology 4](#_Toc164954704)

[3.1 Tools 4](#_Toc164954705)

[3.2 Data Collection 4](#_Toc164954706)

[3.3 Exploratory Data Analysis (EDA) 4](#_Toc164954707)

[3.4 Data Transformation 4](#_Toc164954708)

[3.5 Model Selection 5](#_Toc164954709)

[3.6 API 5](#_Toc164954710)

[4 Results 6](#_Toc164954711)

[4.1 Model Evaluation 6](#_Toc164954712)

[4.1.1 Potential Problems 7](#_Toc164954713)

[4.2 Performance Evaluation 8](#_Toc164954714)

[4.2.1 Full test set 8](#_Toc164954715)

[4.2.2 Single observations 8](#_Toc164954716)

[4.3 What determines the price? 8](#_Toc164954717)

[5 Discussion 9](#_Toc164954718)

[5.1 Data 9](#_Toc164954719)

[5.2 Results 9](#_Toc164954720)

[6 Conclusions 10](#_Toc164954721)

[Appendix A 11](#_Toc164954722)

[References 12](#_Toc164954723)

# Introduction

In everyday life, most people on this earth are exposed to some kind of statistics daily. More often than not this statistic has come from a statistical regression analysis. **“Regression analysis** is a cornerstone technique in statistics and data science that allows us to explore and quantify the relationships between variables. It is used to predict **outcomes**, identify **trends**, and make data-driven **decisions**across various fields, from business and finance to healthcare and engineering.” (*What’s Regression Analysis? A Comprehensive Guide for Beginners*, 2023).

“Statistical learning refers to a set of tools for making sense of complex datasets. In recent years, we have seen a staggering increase in the scale and scope of data collection across virtually all areas of science and industry. As a result, statistical learning has become a critical toolkit for anyone who wishes to understand data — and as more and more of today’s jobs involve data, this means that statistical learning is fast becoming a critical toolkit for everyone.” (James et al., 2023, p.3).

In the plot shown in figure 1 we see data collected using an API from the Swedish statistics agency SCB. The plot shows that newly registered vehicles is in a decline, but the population and demand for passenger vehicles is not.

This tells us that demand for used vehicles is high. Therefore, this report aims to create a regression model that can accurately predict the price of used vehicles, and to find out what variables affects this predicted price. To achieve this objective the following research questions and tasks will be answered.

* Can a model that has a RMSE < 50 000kr be created?
* Can we get an Adjusted R^2 > 0.80 on the same model?
* Which variables effect the price the most?
* Can an API to get data from SCB be created?

A graph of a passenger car

Description automatically generated

Figure 1 API data on new passenger cars

# Theory

## Linear Regression

### Simple Linear Regression

Simple linear regression is a straightforward way for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. To do this, the model predicts the coefficients β0 (intercept) and β1 (slope) by using the least squares fit. The least squares find the best fit by minimizing the residual sum of squares (RSS), Shown in figure 2. (James et al., 2023, p.61-62).

### Multiple Linear Regression

Multiple linear regression works mostly in the same way as simple linear regression. The difference is that in multiple linear regression we extend the model so we can use multiple predictors. We do this by giving each of the predictors a separate slope coefficient in the same model. We still find the best fit by minimizing the residual sum of squares (RSS). (James et al., 2023, p.72).

## Evaluation Metrics

### Residual Sum of Squares (RSS)

RSS is obtained by adding the square of residuals. Residuals are projected deviations from actual data values and represent errors in the [regression](https://www.wallstreetmojo.com/regression/)model’s estimation. A lower RSS indicates that the regression model fits the data well and has minimal data variation. (Pathak, 2024) Shown in figure 2.

A graph of a graph of squares

Description automatically generated with medium confidence

Figure 2 RSS

### Residual Standard Error (RSE)

“The RSE is an estimate of the standard deviation of ϵ". Roughly speaking, it is the average amount that the response will deviate from the true regression line.” (James et al., 2023, p.69).

Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β0 and β1 were known exactly, any prediction would still be of by the RSE amount of units. (James et al., 2023, p.69).

### Adjusted R-squared

Normal R^2 measures the proportion of variability in Y that can be explained. It ranges from 0-1 and a high R^2 is good and a low value is bad. This means that it gives us the proportion on the variance in the Y variable that can be predicted by the predictor variables. The problem with R^2 is that it increases when we add more variables. (James et al., 2023, p.70).

A regression model with many predictor variables has a higher R-squared value, even if the model doesn’t fit the data well. This is where Adjusted R^2 comes in. Adjusted R^2 can tell us how useful a model is and how much of the price that can be explained by the predictor variables, adjusted for the number of predictors in a model*.* (Bobbitt, 2022).

### Bayesian information criterion (BIC)

BIC is a criterion for model selection. It balances the trade off between a models fit and its complexity (number of variables). BIC tends to take on a small value for models with a low-test error and low complexity, so generally we select the model with the lowest BIC value. (James et al., 2023, p.233-234).

### Root Mean Square Error (RMSE)

“One way to assess how well a regression model fits a dataset is to calculate the **root mean square error,** which is a metric that tells us the average distance between the predicted values from the model and the actual values in the dataset. The lower the RMSE, the better a given model is able to “fit” a dataset.” (Bobbitt, 2021).

# Methodology

## Tools

The code was written in R studio. “R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.” (The R Project for Statistical Computing, n.d).

## Data Collection

The data for this report was collected between April 11, 2024, and April 17, 2024. It was collected by a group of students. Each of the group members collected about 500 observations using a web scraper, totalling just under 3 000 observations.

The observations contained information about second hands passenger vehicles for sale on a

Swedish website called Blocket.se. Each observation contained 9 columns: Pris, Bränsle, Växellåda,

Miltal, Modellår, Biltyp, Drivning, Hästkrafter, and Märke. Translated to English: Price, Fuel, Gearbox, Milage, Model year, Type of car, Drivetrain (4WD or 2WD), Horsepower, and Brand.

Each member collected data following the same set of restrictions. This was the filters used:

* Price between 150 000 – 500 000kr.
* Model year between 2014 – 2024.
* Only passenger vehicles.

To minimize the risk of duplicates and to get an even distribution of brands, every member collected their observations from different brands. The brands that were used was Audi, Volvo, Ford, VW, Renault, and Toyota.

## Exploratory Data Analysis (EDA)

The dataset has the dimensions of 2726 rows and 9 columns. There are 4 numeric columns, Pris, Miltal, Modelår, and Hästkrafter. And 5 character columns, Bränsle, Växellåda, Biltyp, Drivning, and Märke.

Bränsle contains: miljöbränsle/hybrid, bensin, diesel, el

Växellåda contains: automat, manuell

Biltyp contains: halvkombi, suv, kombi, sedan, cupé, cab.

Drivning contains: tvåhjulsdriven, fyrhjulsdriven.

Märke contains the brands listed above.

## Data Transformation

The data which was collected contains 5 columns that are non-numeric. For the least squares linear regression model to work there can’t be character data. To solve this, we simply either encode using a dummy variable or change the desired columns to factors. To best suit this report all the character columns was transformed into factors. As seen in figure 3.

A close up of a computer code

Description automatically generated

Figure 3 Factors

## Model Selection

Model selection is an important step during all Machine Learning and regression projects. During this step we must think about the data that will be used, but also what we intend to achieve.

For this report two models were used, one ordinary multiple linear regression model, and one multiple linear regression model with the dependant variable Y transformed with log(Y). The reasoning behind these decisions will be explained further down.

## API

Data was collected from the SCB (Swedish statistics agency) using an API. “An API, or application programming interface, is a set of rules or protocols that enables software applications to communicate with each other to exchange data, features and functionality.” (Goodwin,2024).

To be able to use the API, the library(pxweb) was used in RStudio. To use the API the following line of code was used. Shown in figure 4.

After running this code, we answer the questions in the RStudio console, to specify what data we want. Then after downloading it, the data is ready for use.



Figure 4

# Results

## Model Evaluation

To achieve the objective of this report both evaluation metrics and diagnostic plots was used. Both models were trained on 60% of the whole dataset, evaluated on 20%, and the remaining 20% were used as a test set.

After running both the ordinary multiple linear model called lm\_1 and especially after running the multiple linear model with log(Y) called Log\_lm, the diagnostic plots showed that data points 17, 555, and 61 could be considered outliers, and were therefor removed. Shown in figure 5.

The models were retrained on the new data set. Lm\_1 is now called lm\_3 and Log\_lm is called Log\_lm2. The two different sets of diagnostic plots show that, Log\_lm2s plots look the best. It shows less heteroscedasticity in the Residuals vs Fitted plot, the Normal Q-Q plot shows that the data was more normally distributed, and the Residuals vs Leverage plot shows that the data is more centred around 0. The diagnostic plots can be seen in Appendix A under the names “diagnostic plot lm\_3” and “diagnostic plot Log\_lm2”.

The models were validated on the validation data. The resulting RMSE, Adjusted R^2, and BIC for both models can be seen in figure 6. Lm\_3 got the better RMSE value, but Log\_lm2 got the better Adjusted R^2 and a better BIC, plus the better diagnostic plots. The Log\_ml2 model performs better overall.

A graph of a variety of values

Description automatically generated with medium confidence

Figure 5

A close up of text

Description automatically generated

Figure 6

### Potential Problems

As said by James et al. (2023) “When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:”

#### Non-linearity of the response-predictor relationships

The Residuals vs Fitted plot shows that the data is not strictly linear, but it is linear enough to not be considered a problem. Shown in Appendix A under the name “Diagnostic plot Log\_lm2”.

#### Correlation of error terms

**Durbin-Watson test** is used to detect the presence of autocorrelation in the residuals of a regression. The test has a range from 0 to 4, typically denoted *d*, with *d* = 2 indicating no autocorrelation, *d* < 2 indicates positive serial correlation, and *d* > 2 indicates negative serial correlation. If the *d* is between 1.5 and 2.5 then autocorrelation is likely not a cause for concern. If d is less than 1.5 or greater than 2.5 then there is potentially a serious autocorrelation problem. (Bobbitt, 2021).

Both the residuals plot and the Durbin-Watson test value of 1.952102 show that autocorrelation is

likely not a cause for concern. The Residuals plot can be found in Appendix A.

#### Non-constant Variance of Error Terms

In the diagnostic plot Residuals vs Fitted for Log\_lm2, there was some heteroscedasticity, but not so much that is would be considered a serious problem. The data set this report used had a wide range for the Y variable (Pris), which in it of itself can create some heteroscedasticity. This will be discussed more in the Discussion chapter.

#### Outliers

Outliers has already been mentioned in the model evaluation chapter.

#### High Leverage Points

In the Residuals vs Leverage diagnostic plot, there is a cluster of data points above the leverage value 0.07. After further inspection, all the data points here are cabs. This is most likely nothing wrong with the data points, it is probably a group of sportier cars like a Ford Mustang, so it will not be removed. This will be discussed more in the discussion chapter.

#### Collinearity

“The variance inflation factor (VIF) is the ratio of the variance of βˆj when fitting the full model divided by the variance of βˆj if fit on its own. The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. Typically, in practice there is a small amount of collinearity among the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.” (James et al., 2023, p.102).

The VIF values from the Log\_lm2 model can be seen in figure 7. All the VIF values are < 5. The model is not showing problematic amounts of collinearity.

A number and numbers on a white background

Description automatically generated

Figure 7

## Performance Evaluation

The Log\_lm2 model was the better performer of the two. In the beginning of the code linked with this report, the full data set was split into 3 sets, train, validation, and test. The Log\_lm2 model performance was tested on the test set.

### Full test set

These are the results from the Log\_lm2 models’ evaluation on the test set. Seen in figure 8. The adjusted R^2 and the BIC values are slightly worse for the test set than the validation set. The RMSE value however was slightly better, with about 43 000kr for the validation set and about 39 000kr for the test set. This shows that the Log\_lm2 model is still reasonably good on the test set, as for the validation set.

### Single observations

The model was used to predict 2 observations from the test data separately.

The first observation “car\_data\_test[100, ]” hade the actual price of 357 800kr, and the predicted price for the same observation was 323 851kr. The prediction interval can be seen in figure 9.

The second observation “car\_data\_test[200, ]” had the actual price of 329 800kr, and the predicted price for the same observation was 322 003kr. The prediction interval can be seen in figure 10.

Both these predictions are good, this once again shows that the Log\_lm2 model is a reasonably good performing model.

## What determines the price?

An objective of this report was to understand what and by how much different variables effect the price of a used passenger vehicle.

Milage, Model year, and Horsepower all had \*\*\* significant codes. This means that their respective p-values were <= 0.001, which also means that they are highly significant. Their respective estimates were, Milage = -0.00001699, Model year = 0.06503, and Horsepower = 0.002663. Estimates tells us,

for example if all the other variables were the same, but horsepower increased by on unit (1 hp) the price would go up 0.002663. Estimates for transformed characteristic variables works almost the same as numerical, if for example all the other variables are the same, but the car is an electric vehicle (

which has an estimate = 6.173e-02) the price will change by 0.06173.

As expected Milage, Model year, and Horsepower all were highly significant, but 15 variables in total hade a \*\*\* significance code. 1 variable hade a \*\* significance code and 1 variable hade a \* significance code. All but 2 variables in the “summary (Log\_lm2)” hade a p-value < 0.05, and even if the p-value > 0.05 we can’t say for sure that the variable has no effect on a cars real price. This means that all the variables might affect the price. All the variables and their respective estimates and p-values can be seen in figure Summary in Appendix A.



Figure 8



Figure 9 Figure 10

# Discussion

This project had a somewhat short timeframe, because of this there are some things that could have been done differently.

## Data

A machine learning model is only as good as the data it was trained on.

When the data for this report was collected, a not so restrictive filter was used. This means that data from a wide range of passenger vehicles was collected, this was done so the model could somewhat accurately predict most of the “normal” cars on the Swedish used passenger vehicle market. When using a wide dataset, the prediction accuracy tends to drop. If this project was to be done again, and a better prediction accuracy is wanted. One solution could be to split the data set in to smaller segments with different stricter filters so that the data wouldn’t be as wide. And then train different models on the different data segments. This would likely produce better prediction accuracy, if a car with matching more restrictive filters was predicted using a model with as similar filter.

#### Non-constant Variance of Error Terms and High Leverage Points

Some Non-constant Variance of Error Terms or heteroscedasticity can be seen in the Residuals Vs Fitted diagnostic plot. This could be due to the wide range of the Y variable. A potential solution to this could be the same as mentioned above. If the data set was split into smaller segments with stricter filters, the range of the Y variable would be smaller. This could in turn lead to reduced heteroscedasticity.

The same is true for the high leverage points. In Residuals vs Leverage diagnostic plot, there is a cluster of data points above the leverage value 0.07. This turned out to be a cluster of cabs, probably Ford Mustangs or a similar more expensive sports car. When the data was collected it was known that those types of sports cars could fit trough the filter. The decision to not change the filter was taken, so that many of the “normal” cars with similar variables that was wanted, did not get removed. The high leverage points potential problem could be solved If the sportier cars were to get a data segment and model of their own.

## Results

Keeping in mind the relatively short time frame and the potential problems just discussed, the overall performance of the Log\_lm2 model was satisfactory. To get a RMSE of 39 000kr and an adjusted R^2 0.803 even with the wide range of data and a relatively simple model is an acceptable result.

## Real World Use

If this project were to be completed again and the model’s intended use was to predict price, to buy and sell cars. It would be useful to not only, segment the dataset to smaller sets and train different models, but also to change all the data. The data collected for this project was data from listings of preowned passenger vehicles. To get the best models it would be better to ask Blocket.se to get the data of already sold cars and the actual price the car was sold for. Almost all car listings on Blocket.se intentionally sets the price to high, so the buyer can bargain and buy the car for lower than asking price. This means that even if the model were 100% correct and trained on this projects data, The models still would be wrong compared to the actual value of the car.

# Conclusions

The objective of this report was to create a regression model that can accurately predict the price of used vehicles, and to find out what variables affects this predicted price, with the following questions.

1. Can a model that has a RMSE < 50 000kr be created?
2. Can we get an Adjusted R^2 > 0.80 on the same model?
3. Which variables effect the price the most?
4. Can an API to get data from SCB be created?

Answers:

1. Yes, the Log\_lm2 model had a RMSE of 43 000kr on the validation set and a RMSE of 39 000kr on the test set, both of which are < 50 000kr.
2. Yes, the Log\_lm2 model got an Adjusted R^2 of 0.84 on the validation set and an Adjusted R^2 of 0.803 on the test set, both of which are > 0.8.
3. Almost all the variables had a statistically significant effect on the price. The variables with a p-value < 0.001 were, Bränslediesel, Bränsleel, Växellådamanuell, Miltal, Modellår, Biltyphalvkombi, Biltypkombi, Biltypsedan, Biltypsuv, Drivningtvåhjulsdriven, Hästkrafter, Märkeford, Märkerenault, and Märkevolkswagen. Translated to English they are called, Diesel, Electric, Manual, Milage, Model year, Hatchback, Station Wagen, Sedan, SUV, Two-wheel drive,

Horsepower, Ford, Renault, Volkswagen.

1. Yes, an API to collect data from SCB was created using the line of code shown earlier in this report. Shown in figure 4.

# Appendix A

Link to all the code in Github:

A group of graphs showing different values

Description automatically generated A collage of graphs and diagrams

Description automatically generated

Diagnostic plot lm\_3 Diagnostic plot Log\_lm2

A graph with numbers and dots

Description automatically generated A screenshot of a computer

Description automatically generated

Residuals plot Summary

# References

Bobbitt. (2022, March 24). *How to Interpret Adjusted R-Squared*. STATOLOGY. Retrieved April 24, 2024, from https://www.statology.org/adjusted-r-squared-interpretation/

Bobbitt. (2021, May 10). *How to Interpret Root Mean Square*. STATOLOGY. Retrieved April 24, 2024, from https://www.statology.org/how-to-interpret-rmse/

Goodwin. (2024, April 9). *What is an API?* IBM. Retrieved April 24, 2024, from https://www.ibm.com/topics/api

James, Witten, Hastie , & Tibshirani. (2023, June 21). *An Introduction to Statistical Learning with Applications in R* (Second Edition). Springer.

Pathak. (2024, April 4). *Residual sum of squares*. WallStreetMojo. Retrieved April 23, 2024, from https://www.wallstreetmojo.com/residual-sum-of-squares/

*The R Project for Statistical Computing*. (n.d.). R-project. Retrieved April 24, 2024, from https://www.r-project.org/

*What’s Regression Analysis? A Comprehensive Guide for Beginners*. (2023, May 1). LEARN STATISTICS EASILY. Retrieved April 23, 2024, from https://statisticseasily.com/whats-regression-analysis/