Price prediction of cars using

Multiple Linear Regression



Gayathree Basarahalli

EC Utbildning

R

202504

# Abstract

This project aims to predict the prices of used Volvo cars using multiple linear regression with stepwise selection. The data was collected manually from Blocket with the help of all classmates. The focus was on cars from 2014 to 2025, as it was assumed that cars manufactured after 2014 adhere to Euro 6 standards, which will increase model accuracy. Multiple linear regression model with stepwise selection was fitted to identify the most relevant features and predict prices of cars.

**Keywords:** Blocket, Used Volvo cars, Multiple linear regression, Stepwise selection.

Table of contents

[Abstract 2](#_Toc196518570)

[1 Introduction 4](#_Toc196518571)

[2 Theory 4](#_Toc196518572)

[2.1 Multiple Linear Regression (MLR) 5](#_Toc196518573)

[2.2 Stepwise selection 6](#_Toc196518574)

[2.2.1 Forward Stepwise Selection: 6](#_Toc196518575)

[2.2.2 Backward Stepwise Selection: 6](#_Toc196518576)

[2.2.3 Hybrid Stepwise Selection: 7](#_Toc196518577)

[3 Method 8](#_Toc196518578)

[3.1 Import necessary Libraries: 8](#_Toc196518579)

[3.2 Data Preprocessing : 8](#_Toc196518580)

[3.3 Multiple linear regression 10](#_Toc196518581)

[3.4 Evaluation with Test dataset and Validation dataset 10](#_Toc196518582)

[4 Result and Discussion 11](#_Toc196518583)

[4.1 RMSE and R-Square test 11](#_Toc196518584)

[4.2 Prediction and Confidence interval 11](#_Toc196518585)

[5 Conclusions 12](#_Toc196518586)

[References 13](#_Toc196518587)

# Introduction

The current report considers data collected manually from Blocket, a Swedish online marketplace where individuals and companies buy and sell their products. The main reason to consider this data is to be relevant to current market and ensuring the model to be built on real -world data. To be relevant to Swedish market, collected used Volvo cars data from all the regions in Sweden.

In this report, Multiple linear regression model is used to predict prices. Step wise selection, both forward and backward is applied to identify most contributing features for price prediction.

Relatable questions to be considered to achieve the aim of the report:

* Develop Multiple regression model with greater accuracy?
* Does applying a log transformation to the target variable improve the model's performance
* Major features influencing price prediction?
* How does stepwise regression help model to perform better?

# Theory

## Multiple Linear Regression (MLR)

**Multiple Linear Regression (MLR)** is an extension of simple linear regression that allows us to predict a response variable using more than one predictor. While simple linear regression uses only a single explanatory variable to model the relationship with the response, MLR includes multiple predictors in a single unified model.

In practice, this is important because many real-world outcomes are influenced by several factors simultaneously. For example, in the case of advertising data, sales may be influenced not just by TV ads, but also by spending on radio and newspapers. Fitting separate simple regression models for each predictor can lead to misleading results, especially if the predictors are correlated with each other. In such cases, the effect of one variable can be confounded by another.

To address this, MLR combines all predictors into one model:

Y=β0​+β1​X1​+β2​X2​+⋯+βp​Xp​+ϵ

Here:

* Y is dependant or target variable.
* X1,X2...Xp are features or predictors.
* β0 is the intercept
* β1​,β2​,....βp​ are the coefficients that measure the impact of each predictor on the response, holding all other predictors constant.
* ϵ is the error term capturing the variation not explained by the model.

This approach allows for a more accurate and interpretable model, especially when features influence target. It also helps avoid issues that come from fitting multiple separate models, such as inconsistent predictions or overestimating effects.

In this project, Multiple Linear Regression is used to predict used Volvo car prices using features like year, mileage, fuel type, and others. Stepwise selection is applied to identify the most important predictors.

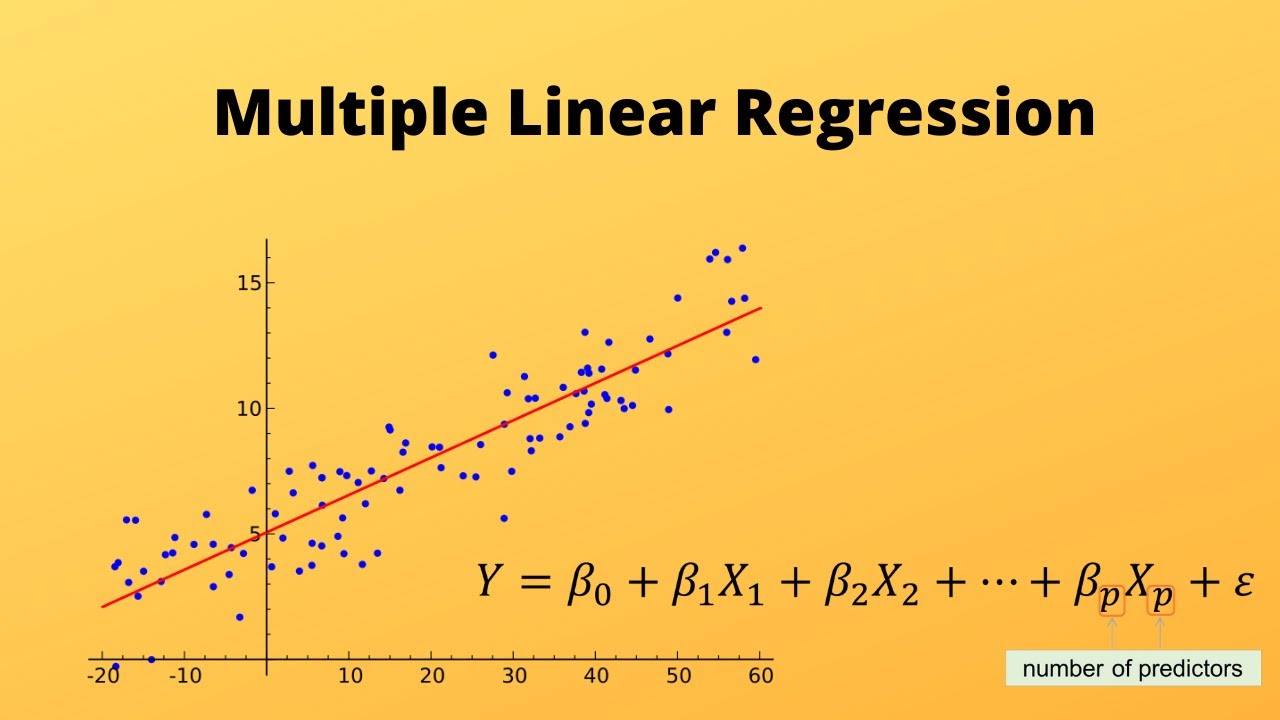


Figure 1 Representation of multiple linear regression

## Stepwise selection

Stepwise selection is a method used for model selection that is more computationally efficient and statistically robust than best subset selection, especially when the number of predictors (**p**) is large. It explores a smaller subset of models and helps reduce the risk of overfitting and high variance in coefficient estimates.

### Forward Stepwise Selection:

* Starts with the null model (no predictors).
* At each step, adds one predictor that improves the model the most (based on smallest RSS or highest R²).
* Continues adding until all predictors are included.
* In total, it fits 1 + p(p + 1)/2 models, which is much fewer than the 2^p models in best subset selection.
* The best model is selected from all the models using criteria like validation error, Cp (AIC), BIC, adjusted R², or cross-validation.
* **Limitation**: May not find the globally best model if a better two-variable model does not include the best one-variable predictor.

### Backward Stepwise Selection:

* Starts with the full model (all p predictors).
* At each step, removes one predictor that contributes the least.
* Continues removing until no predictors remain.
* Also fits 1 + p(p + 1)/2 models.
* The best model is selected using similar criteria as in forward selection.
* Limitation: Cannot be used when p ≥ n because the full model cannot be fit.

### Hybrid Stepwise Selection:

* Combines forward and backward selection.
* Adds predictors one-by-one, like forward selection.
* Removes predictors if they no longer improve the fit, mimicking best subset selection.
* Retains computational efficiency while improving flexibility.

# Method

## Import necessary Libraries:

To work with this model, we need to import all necessary libraries.

library(ggcorrplot)

library(lubridate)

library(caret)

library(glmnet)

library(tidyverse)

library(Amelia)

library(car)

## Data Preprocessing :

The data was collected manually from Blocket, platform where individuals and companies sell and buy products. Collectively, our classmates divided work and collected used Volvo cars data from different regions in Sweden. Primarily, the data set consists of 615 observations with 16 variables.

Data set is cleaned with null values and non-numerical values, after processing data contains 605 rows. Not all variables are influencing price prediction so removed some variables for better handling. Finally, the dataset contains 605 rows with 12 variables. Converting all categorical values to factors which help in dummy encoding.

Data is divided into 3 datasets: 70% of training data, 15 % of each validation and test data.

Exploratory Data Analysis (EDA) was performed, which included examining the data structure and visualizing the plots. From fig(2) data is right skewed and requires log transformation for target variable.

A graph of selling price

AI-generated content may be incorrect.

Figure 2 Selling price – shows right skewness

A graph with red and blue squares

AI-generated content may be incorrect.

Figure 3 Correlation matrix – relationship among numerical variables

## Multiple linear regression

In this model, all the variables are included while fitting the model. Apply step wise selection in both directions forward and backward, helps in selecting highly influential variables to fit into the model for better accuracy and model development.

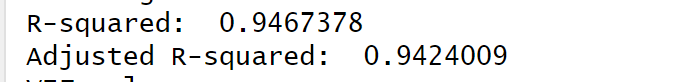
Align factors for ensuring consistency in factor levels between train data, test and validation data.

## Evaluation with Test dataset and Validation dataset

Evaluating the model by using RMSE and R-square test score. Vif is calculated to check multicollinearity. These measures show predictive performance.

# Result and Discussion

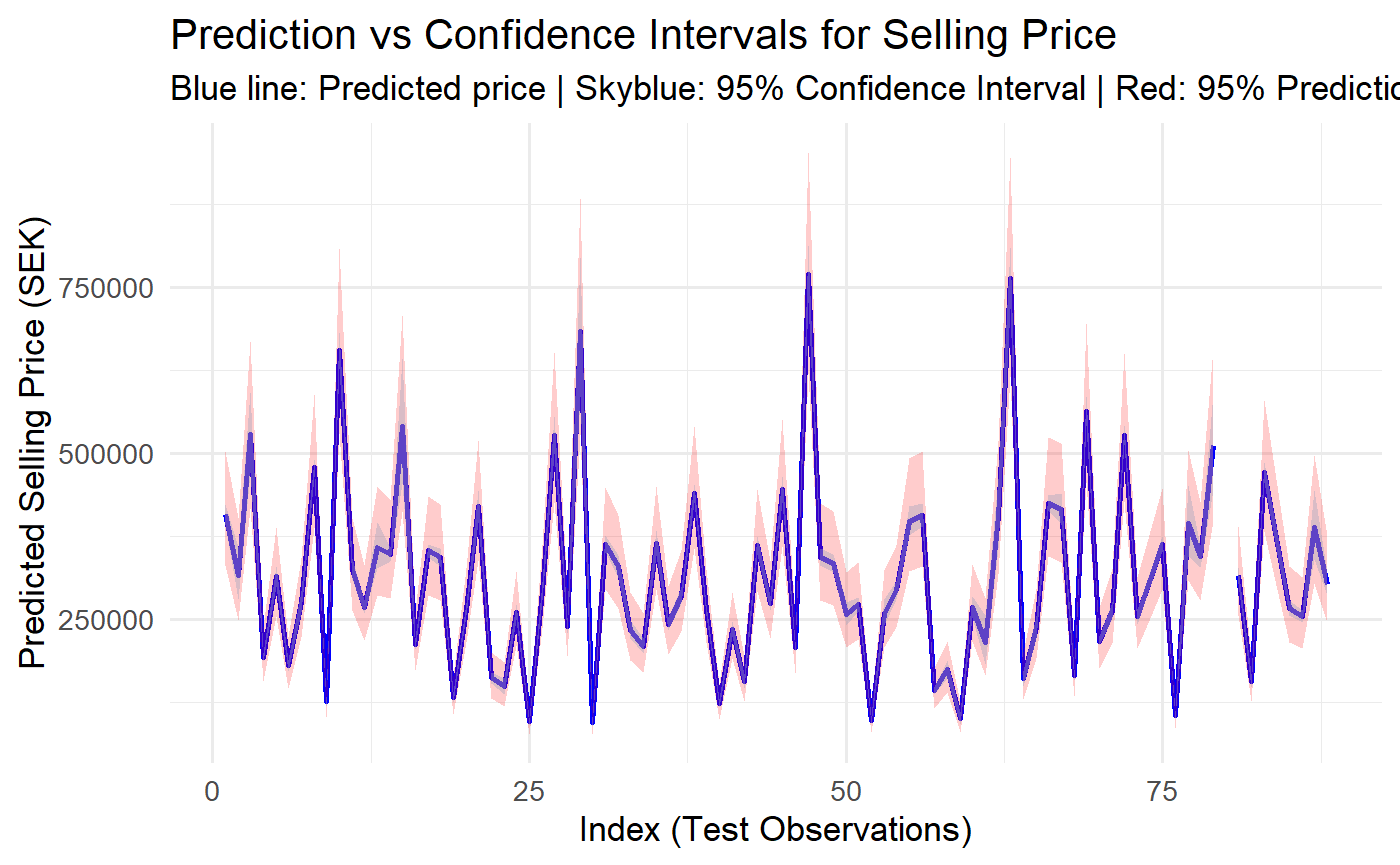
## RMSE and R-Square test



*Figure 4: Final values of the model*

## Prediction and Confidence interval

Confidence interval reflects uncertainty around the predicted price, while the wider prediction intervals represent where the car prices are expected to fall. Form the below figure, fitted model provide accurate values for unseen data.



*Figure 4: Confidence interval and Prediction interval*

# Conclusions

The following conclusions regarding setting out objectives we considered in the beginning:

* After preprocessing, the model was initially fit with all variables. Using stepwise selection, a more accurate model was achieved with an R-squared of around 94%, indicating improved performance with fewer, more relevant predictors.
* Log transformation helped in making target variable values more normally distributed.
* Based on the model, Model year, Mileage, Fuel, Car type are majorly influencing price prediction.
* Step wise selection helps in automatically selecting variables one after other and exclude less influential variables.

# References

James, G., Hastie, T., Witten, D., & Tibshirani, R. (2021). An introduction to statistical learning

Prgomet, A. (2023, April 5). R Programming [Video]. YouTube.

<https://www.youtube.com/watch?v=NcxMuCG6FS8&list=PLgzaMbMPEHEyLy3NJ8tZqHBzoVcZlowX4&index=2>