Regression analysis

of manually collected data on cars for sale

on the Blocket website



Lidiia Kashevarova

EC Utbildning

Kunskapskontroll, R

202404

# Abstract

This study involves both group and individual work. The group, consisting of 6 students, compiled a database comprising 352 observations gathered from advertisements for the sale of Volvo V60 automobiles.

During individual work, a linear regression was developed, enabling the assessment of data quality, confirmation of the linear relationship between independent and dependent variables, determination of the significance of various parameters, and prediction of car prices with R-squared - 0.8469207 (verified on test data).

Additionally, data was obtained from the Central Statistical Office (CSO) using the PX-WEB API; which made it possible to identify a trend toward an increase in the share of registrations of more environmentally friendly electric vehicles or hybrids.

# Table of contents

Abstract ..................................................................................................................................................2 Introduction........................................................................................................................................... 4 1.Theory................................................................................................................................................ .5

1.1.Linear regression………………………………………………………………………………………………………………………5

1.2. Assessing the Accuracy of the Model………………………………………………………………………………………5

1.2.1. Residual Sum of Squares (RSS)……………………………………………………………………………………….5

1.2.2. Residual Standard Error ………………………………………………………………………………………………..6

1.2.3. R^2 Statistic ………………………………………………………………………………………………………………….6

1.2.4. Adjusted R-Squared………………………………………………………………………………………………………6

1.3. Significance of Variables………………………………………………………………………………………………………..6

1.4. Variance Infation Factor (VIF)………………………………………………………………………………………………..7

2. Method ………………………………………………………………………………………………………………………………………..8

3. Data Collection………………………………………………………………………………………………………………………………9

4. Resultat och Diskussion…………………………………………………………………………………………………………………10

4.1. Model 1………………………………………………………………………………………………………………………….10

4.2. Model 2………………………………………………………………………………………………………………………….12

4.3. Validating models on the validation dataset…………………………………………………………………..12

4.4. Model 1 testing……………………………………………………………………………………………………………..13

4.5. The data from the Central Statistical Office (SCB)………………………………………………………….14

5. Conclusion……………………………………………………………………………………………………………………………………17

6. Theoretical issues…………………………………………………………………………………………………………………………19

7. Self-evaluation……………………………………………………………………………………………………………………………..21

8. Referenser ……………………………………………………………………………………………………………………………………23

# Introduction

“R provides a wide range of statistical (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, etc.) and graphical methods and is highly extensible.” (The R Project for Statistical Computing, Internet) The R statistical software package allows you to implement statistical learning methods. “Broadly defined, supervised statistical learning involves building a statistical model to predict or estimate outcomes based on one or more inputs.” (G.James &D.Witten &T.Hastie &R.Tibshirani , p. 1) Therefore, using statistical learning and the statistical software package R, we can draw conclusions about the significance of variables and their relationships, predict results and evaluate the accuracy of forecasts, as well as visualize the data and results.

The practical application of statistical learning spans various domains. Statistical learning allows us to identify data that significantly impacts outcomes while excluding factors with minimal effects. Such an approach aids decision-making and process optimization, potentially leading to substantial cost reduction and increased productivity. Accordingly, this opens up wide opportunities for the use of statistical learning and the R statistical software package in business, production, medicine and other aspects of our lives.

This study, using the example of the Volvo V60 car, provides insight into which characteristics of the vehicle have the greatest influence on price formation. The model developed during the study accurately predicts the price of the Volvo V60 car.

# Teori

1.1.Linear regression

Despite its simplicity, linear regression continues to be a useful tool for predicting a quantitative response and determining the strength of the relationship between variables and the influence of independent variables on the dependent variable .

𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀

Y - dependent variable

𝑥1, 𝑥2…𝑥𝑝 - independent variables

𝛽0 *- the* intercept

 𝛽1,  𝛽2… 𝛽p -slope, they show how the dependent variable changes if the corresponding independent variable changes by 1 unit and all other explanatory variables are held constant

𝜀 - diference between the observed response value and the response value that is predicted by linear model 𝜀 = yi −yˆi (G.James &D.Witten &T.Hastie &R.Tibshirani , p.82-88)

1.2. Assessing the Accuracy of the Model

1.2.1. Residual Sum of Squares (RSS)

The residual sum of squares (RSS) is used to account for the accuracy of the model. It serves as an indicator of how well the model fits the data. It is calculated as the sum of the squares of the differences between the actual values of the dependent variable and the values predicted by the model:

RSS = (𝜀1)^2 + (𝜀2)^2 + ··· + (𝜀n)^2 (p.62)

or

En bild som visar Teckensnitt, text, vit, design

Automatiskt genererad beskrivning

If the RSE is quite large, it may indicate that the model is not processing the data well (G.James &D.Witten &T.Hastie &R.Tibshirani , p.69)

1.2.2. Residual Standard Error

RSE provides an absolute measure of lack of ft of the model

En bild som visar text, Teckensnitt, vit, linje

Automatiskt genererad beskrivning

”If yˆi ≈ yi for i = 1,...,n—will be small, we can conclude that the model fts the data very well” (G.James &D.Witten &T.Hastie &R.Tibshirani , p.69)

We can calculate the percentage error as: (RSE/mean value of Y)\*100%.

1.2.3. R^2 Statistic

RSE provides an absolute measure of the fit of a model to the data. But since it is measured in Y units, it is not always clear what constitutes a good RSE. R2 statistics provide an alternative RSE measure. It takes the form of a proportion a fraction of the variance and takes a value between 0 and 1 regardless of the scale of Y (G.James &D.Witten &T.Hastie &R.Tibshirani , p.69)

En bild som visar text, Teckensnitt, linje, vit

Automatiskt genererad beskrivning

TSS = (yi − y¯)2 - the total sum of squares

” An R2 statistic that is close to 1 indicates that a large proportion of the variability in the response is explained by the regression. A number near 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance σ2 is high, or both.” (G.James &D.Witten &T.Hastie &R.Tibshirani , p.70)

1.2.4. Adjusted R-Squared

”Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.” (Investopedia, Internet)

1.3. Significance of Variables

T-value is used to test the significance of the coefficient. The greater the absolute value of this value, the more significant the coefficient.

Pr(>|t|) is the probability that t will have a greater absolute value if a certain coefficient is 0.

En bild som visar text, skärmbild, Teckensnitt, vit

Automatiskt genererad beskrivning

( G.James &D.Witten &T.Hastie &R.Tibshirani , p.67)

The lower the value of Pr(>|t|), the more significant the coefficient:

* very high innovativeness (Pr(>|t|) < 0.001);
* high innovativeness (0.001 < Pr(>|t|) < 0.01);
* moderate innovativeness (0.01 < Pr(>|t|) < 0.05 ));
* insignificant coefficients (Pr(>|t|) > 0.1).

F-statistic and its corresponding p-value help evaluate the overall suitability of the model. They help to estimate effect of all the variables together. The smaller the p-value, the greater the significance of variables.

1.4.Variance Infation Factor (VIF)

One potential problem of a regression model could be collinearity or multicollinearity (collinearity between three or more variables). "A better way to assess multicollinearity is to compute the variance inflation factor (VIF).

En bild som visar text, Teckensnitt, vit, nummer

Automatiskt genererad beskrivning

(RXj |X−j)^2 is the R2 from a regression of Xj onto all of the other predictors. If (RXj |X−j)^2 is close to one, then collinearity is present, and so the VIF will be large.” ( G.James &D.Witten &T.Hastie &R.Tibshirani , p.103)

# Method

For the regression analysis, a database of cars listed for sale on the Blocket website was utilized. A total of 352 advertisements for the sale of cars were selected.

Given the wide price range of cars, a Volvo V60 model was chosen for analysis. The selection of this model was based on several criteria:

- it is widely available in Sweden, with over 1500 listings for sale on the website spanning with model years 2018 to 2023;

- it belongs to the mid-range class of models, with price ranges comparable to other popular models such as the Volkswagen Caddy, BMW 118, Kia Sportage Plug, and Audi A4 of the same years;

- it offers various configurations (Fuel: Hybrid, Gasoline, Diesel; Transmission: Automatic, Manual; varying Horsepower), all of which may influence the car's price.

Thus, the choice of a single car brand enabled the exclusion of the influence of immeasurable factors such as prestige or reputation, focusing instead on the impact of measurable characteristics of the car itself.

The dependent variable chosen was the price, while the independent variables included:

* Model Year: Categorical variable with levels 2018, 2019, 2020, 2021, 2022, and 2023;
* Mileage (continuous variable);
* Transmission: Categorical variable with levels Automatic (0) and Manual (1);
* Fuel Type: Categorical variable with levels Hybrid/Electric (0), Gasoline (1), and Diesel (2);
* Horsepower (continuous variable).

The database was divided into a training set of 60%, a validation set of 20%, and a test set of 20%.

In RStudio, linear regressions were constructed as follows:

* a regression model using all independent variables;
* a regression model using all independent variables except Gearbox.

Following testing on the test data, another dataset consisting of 20 rows with various models was created (Volkswagen Caddy, BMW 118, Kia Sportage Plug, and Audi A4). The criteria for selecting data from advertisements were that the price of models from a specific year falls within the price range of the Volvo V60 of that year.

In addition, the work used SCB Statistikmyndigheten Statistikdatabasen/Transporter och communicator/Fordonsstatistik/Fordonsstatistik/Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2024M03 for the period from January 2018 to December 2023 (SCB, Internet)

# Data Collection

The group included the following students: Anders, Jacob, Manna, Aikaterini, Girlie and Lidiia.

At our initial meeting, we decided to compile a database for the Volvo V60. Each student took responsibility for a specific model year, thus avoiding duplication of advertisements:

*Table 1. Data collection team members*

|  |  |
| --- | --- |
| Name | Model year |
| Anders | 2018 |
| Jacob | 2019 |
| Manna | 2020 |
| Aikaterini | 2021 |
| Lidiia | 2022 |
| Girlie | 2023 |

We also discussed what parameters could theoretically influence the cost of a car and decided what to include in our database: mileage, transmission, fuel type and horsepower, price. Transmission and Fuel Type are in text format, so we have defined these variables to be categorical.

We created a document in Google Doc where each student entered their data. As a result, we got a database consisting of 352 rows.

Most of the data was collected on the first day so I was able to get started pretty quickly. It was helpful to discuss what data we should collect.

I believe that we did everything as efficiently as possible, so I would not change anything.

# Resultat och Diskussion

4.1. Model 1

The first model is a linear regression and includes all independent variables. The dependent variable is the cost of the cars.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

From the summary we can draw the following conclusions:

1. The variable "mil" has a negative effect on the price of cars, which is natural as the price of the car decreases with an increase in mileage. Similarly, the variable "Gearbox\_Automat\_0\_Manuell\_1" also has a negative effect, which is explainable since in the data table, "Automat" is denoted as 0 and "Manuell" is denoted as 1. Consequently, cars with a manual gearbox are cheaper than cars with an automatic gearbox.
2. The expected outcome does not match the fact that the variable

Fuel\_Hybrid\_0\_Bensin\_1\_Diesel\_2 is positive. Since the hybrid in the table has a value of 0 and diesel has a value of 2, it was expected that an increase in the value of Fuel\_Hybrid\_0\_Bensin\_1\_Diesel\_2 would lead to a decrease in the car's price.

1. The intercept, Model\_year, mil, and Horsepower\_Hp exhibit very high significance. The variable Fuel\_Hybrid\_0\_Bensin\_1\_Diesel\_2 demonstrates high significance. The variable Gearbox\_Automat\_0\_Manuell\_1 shows moderate significance.
2. Residual standard error: (RSE/mean value of Y)\*100%

(35280/315575)\*100%=11,17%

We can conclude that the model fits the data very well.

1. R-square value close to 0.8315 indicates that most of the variability in the response is

esplanad by regression. .

1. The low p-value for the F-statistic provides evidence to support the claim that all variables are significant and the model is adequate
2. From the obtained VIF data we can see that there is a relationship between the corresponding variable and the other variables in the model. However, the value does not exceed 3.

En bild som visar text, Teckensnitt, skärmbild, vit

Automatiskt genererad beskrivning

Let's also consider the plots of the first linear regression model.

En bild som visar text, diagram, skärmbild, linje

Automatiskt genererad beskrivning

*Figure 1: The plots of the first linear regression model*

1. The "Residuals vs Fitted" plot illustrates the differences between the actual values and the values predicted by the model. If there is a linear relationship between the independent variables and the dependent variable, the points should be scattered in such a way that no pattern can be discerned. In this case, we can assume a linear relationship and lack of correlation among the residuals.
2. The Q-Q Plot is used to check the normality of the residuals distribution. If the points follow a straight line, it indicates a normal distribution of residuals.
3. The Scale-Location Plot helps to check the constancy of dispersion of residuals (homoscedasticity) . In our case, the points are distributed quite randomly. It is difficult to judge the presence of homoscedasticity from the plot.
4. Residuals vs Leverage helps identify observations that have a significant influence on the regression model.

Overall, the distribution of most points on the plots corresponds to the expected outcome and does not indicate any deviations. However, there are several points that are far away. This may indicate an anomalous value or an outlier. Let's examine them:

* row 78: A 2019 model priced at 370,000, significantly higher than the prices of other 2019 cars;
* row 105: Here, it is more difficult to observe anomaly. I assume that a combination of all variables plays a role here. Cars with almost identical characteristics are priced 25,000-30,000 cheaper.

The anomaly in lines 196 and 48 is difficult to assess at first glance. I think a detailed comparison of all the characteristics is necessary here.

4.2. Model 2

Model 3 also constitutes a linear regression. In Model 3, I emitted the variable Horsepower\_Hp which has highest VIF value.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

From the summary we can draw the following conclusions that all independent variables are very high significance however, RSE increased R-squared decreased.

En bild som visar text, diagram, skärmbild, linje

Automatiskt genererad beskrivning

*Figure 2: The plots of the second linear regression model*

In the plots, no significant changes are observed compared to the first model

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

Now all variables have VIF below 2.

4.3. Validating models on the validation dataset

En bild som visar text, Teckensnitt, skärmbild, linje

Automatiskt genererad beskrivning

The validation results indicate that Model 1 outperforms Model 2 in terms of RMSE, Adj\_R\_squared, and BIC values.

4.4. Model 1 testing

When testing the model on test data which contains only Volvo V 60, the following results were obtained:

* RMSE - 39774.99
* R-squared - 0.8469207

As a result of the fact that the database contains data only from the Volvo V60, the model fits the data perfectly and shows high accuracy. The model can be used to predict the cost Volvo V60. For validation parameters of three cars were entered and their prices were forecasted using Model 1. On the Blocket website we can see it is observed that the majority of cars fall within the prediction intervals.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning

To assess how the model performs with other car brands, another database was created containing cars whose prices fall within the price range of a Volvo V60 of the same year. Model 1 showed the following results:

* RMSE 60172.19
* R- squared 0.5842901

From the data obtained it is clear that the performance of the model has decreased significantly.

4.5. The data from the Central Statistical Office (SCB)

According to the results of the regression analysis, it was found that the type of fuel used in the car is highly significant for the car's price. However, contrary to expectations, for example, diesel increases the car's price compared to gasoline or hybrid. This fact, along with the relevance of the issue of transitioning to more environmentally friendly types of fuel, prompted me to investigate how the overall number of more environmentally friendly cars is changing in Sweden. For this purpose, I used data from the Central Statistical Office (SCB) for the period from January 2018 to December 2023, from the table "Nyregistrerade personbilar efter region, drivmedel och månad" (SCB, Internet) . I compared the number of car registrations using gasoline and diesel with hybrid or electric cars.

*Table 2. Statistical information on the number of registered cars with different types of fuel*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2018** | **2019** | **2020** | **2021** | **2022** | **2023** |
| **bensin** | 173808 | 169198 | 114640 | 97471 | 67247 | 64777 |
| **diesel** | 137409 | 117775 | 68073 | 53734 | 36788 | 26498 |
| **el** | 7147 | 15795 | 28097 | 57881 | 96163 | 112775 |
| **elhybrid** | 21023 | 33123 | 22631 | 24137 | 27584 | 26540 |
| **bensin+diesel** | 311217 | 286973 | 182713 | 151205 | 104035 | 91275 |
| **el+elhybrid** | 28170 | 48918 | 50728 | 82018 | 123747 | 139315 |
| **procent of bensin+diesel** | 92 | 85 | 78 | 65 | 46 | 40 |
| **procent of el+elhybrid** | 8 | 15 | 22 | 35 | 54 | 60 |

Next, I compiled a table from data about Volvo V60 cars collected manually on the Blocket website

*Table 3. Cars Volvo V60 collected manually on the Blocket website*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **2018** | **2019** | **2020** | **2021** | **2022** | **2023** |
| **bensin** | 11 | 13 | 5 | 12 | 7 | 26 |
| **diesel** | 28 | 33 | 36 | 25 | 33 | 30 |
| **elhybrid** | 15 | 4 | 8 | 33 | 20 | 13 |
| **bensin+diesel** | 39 | 46 | 41 | 37 | 40 | 56 |
| **procent of bensin+diesel** | 72 | 92 | 84 | 53 | 67 | 81 |
| **procent of elhybrid** | 28 | 8 | 16 | 47 | 33 | 19 |

*Figure 2: The proportion of cars with diesel or gasoline engines*

*Figure 3: The proportion of cars with el or elhybrid engines*

Based on the above data, we can observe a clear trend in Sweden of decreasing registrations of gasoline or diesel cars and increasing registrations of electric cars and hybrids. However, the data on Volvo V60, which were collected manually, do not show such a trend, and overall, there are significantly more diesel and gasoline cars. For example, in 2023, 40% of cars were registered with diesel and gasoline fuels in Sweden as a whole, while in our Volvo V60 dataset, such cars accounted for 81%. This may be due to the fact that the Volvo V60 model is more commonly equipped with diesel or gasoline fuel types by design or hybrid owners may be less inclined to sell their cars on the secondary market.

# Conclusion

The following conclusions regarding set out objectives, tasks and research questions can be drawn:

1. Working in a group is a good solution when it is necessary to collect a large amount of data manually. However, it is important to discuss all the details, clearly distribute the areas of work, and agree on deadlines.

2. Using only one model, the Volvo V60, in the analysis helped avoid many potential problems associated with data quality and focus on a detailed examination of factors influencing the formation of the car's price.

3. The analysis conducted using linear regression showed that the data are normally distributed and do not contain a large number of outliers. All collected characteristics have a certain degree of significance:

- model year positively influences the car's price and has a very high level of significance;

- mileage negatively influences the car's price and has a very high level of significance;

- horsepower positively influences the car's price and has a very high level of significance;

- automatic transmission increases the car's price compared to manual transmission, but overall, this factor has moderate significance;

- there are some doubts about the fuel type used in the car, as the analysis showed that diesel increases the car's price compared to hybrid.

4. When testing the models on test data, the following results were obtained: RMSE (Root Mean Square Error) of 39774.99 and an R-squared value of 0.8469207. The model was even tested on other car brands (not Volvo V60), resulting in an RMSE of 60172.19 and an R-squared value of 0.5842901. For validation parameters of three cars were entered and their prices were forecasted using Model 1. On the Blocket website we can see it is observed that the majority of cars fall within the prediction intervals.

5. From all of the above, we can conclude that linear regression is quite suitable for analyzing prices for cars in the secondary market. The independent variables for linear regression were selected correctly. A database of 350 observations is sufficient for regression analysis and obtains a high level of prediction accuracy and model fit to the data for one car model.

6. To understand the general trend of the ratio of more environmentally friendly cars (electric or hybrid) to less environmentally friendly ones (diesel and gasoline), additional data analysis was conducted using the Central Statistical Office (SCB) website. For this analysis, data was obtained using the PX-WEB API. This analysis revealed a clear trend of decreasing registrations of diesel and gasoline cars in Sweden from January 2018 to December 2023. However, the manually collected data from the Blocket website on the sale of Volvo V60 cars do not show the same trend. From this, it can be concluded that the fuel type indicator in the model is ambiguous.

6. Theoretical issues

1. Kolla på följande video: https://www.youtube.com/watch?v=X9\_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

Quantile-Quantile (QQ) plot är en metod som används för att undersöka antagandet att data är normalfördelade. Om en uppsättning observationer är approximativt normalfördelad, kommer en Quantile-Quantile (QQ) plot att vara ungefär en rät linje.

 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?

Karin har rätt.

I maskininlärning ägnar vi särskild uppmärksamhet åt att bygga modeller och deras prediktionsnoggrannhet. Vi tränar modellen på tillgänglig data så att modellen senare kan göra förutsägelser om data som ännu inte är kända.

Å andra sidan involverar statistisk regressionsanalys ofta både förutsägelse och statistisk slutledning. Till exempel testar vi hypotesen att någon parameter är lika med noll eller kollar vilka parametrar som har större inverkan på den beroende variabeln.

 3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

Konfidensintervall baseras på  snittvärden av en observation för en viss grupp.

 𝑌 = 𝛽0 + 𝛽1\**X*

Prediktionsintervall baseras på  värde för en individ och inkluderar osäkerhet epsilon

 𝑌 = 𝛽0 + 𝛽1\**X* + 𝜀

4. Den multipla linjära regressionsmodellen kan skrivas som:

𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 . Hur tolkas beta parametrarna?

𝛽0 *- interceptet,*

 𝛽1,  𝛽2… 𝛽p -lutningen, de visar hur förändrar beroende variabeln om den motsvarande  oberoende variabeln ändras med 1 enhet  och alla andra förklaringsvariabler hålls konstanta

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

Nej, vi behöver inte använda validering eller test set eftersom måtten BIC uppskattar test error från training error.

6. Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

1. Välj en nollmodell M0 som inte innehåller predictors (x) och förutsäger enkelt medelvärdet för varje observation

2.  Vi ökar monotont antalet funktioner som ingår i modellerna (k=1, k=2 ....k=p) och definierar linjära regressionsmodeller för varje möjlig uppsättning prediktorer och väljer den bästa modellen Mk (modellen med minsta RSS eller största R^2).  Låg RSS eller hög R2 indikerar en modell med lågt träningsfel men vi vill ha en modell med lågt testfel.

3. I det tredje steget måste vi välja modellen med det minsta testfelet. För detta kan vi använda Cp (AIC), BIC eller justerad R2. Eller använd korsvalideringsmetoden. Som ett resultat väljer vi en bästa modell bland M0,...,Mp.

Ett potentiellt problem med att ”Best Subset Selection”  är beräkningsbegränsningar. Antalet möjliga modeller att överväga växer snabbt när p ökar. “Consequently, best subset selection becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers” (p.228)

 7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

Modeller kan inte vara 100 procent realistiska eftersom de är mycket förenklade men trots det kan modeller ge oss värdefull information.

# Självutvärdering

1. Challenges you had during work and how you handled them.

Model selection. I tried Ridge and Lasso regression, but their results were quite difficult to interpret. In the end, I decided to exclude them, since the linear model gives quite good results.

Get data from the Central Statistical Bureau (SCB) website using the API. I tried to find an answer on the Internet, asked a question in the GPT chat

1. What grade do you think you should have and why.

G or VG

# Referenser

1. G.James &D.Witten &T.Hastie &R.Tibshirani, An Introduction to Statistical Learning with Applications in R (p.1-88)
2. The R Project for Statistical Computing , <https://www.r-project.org/about.html> (16.04.2024)
3. Investopedia, <https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp#:~:text=Adjusted%20R%2Dsquared%20is%20a,model%20by%20less%20than%20expected>. (16.04.2024)
4. Statistics How To, <https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/> (16.04.2024)
5. SCB Statistikdatabasen, [Nyregistrerade personbilar efter län och kommun samt drivmedel. Månad 2006M01 - 2024M03. PxWeb (scb.se)](https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/) (16.04.2024)