Regression Analysis

Investigating the Effect of Average Income Level on Car Prices



Melike Koyuncu  
EC Utbildning

R Programmering

202404

# Abstract

This project investigates the relationship between car prices and average income levels using regression analysis in R. Data from multiple car brands and models was collected, including variables such as car price, average income level, and other relevant factors. The primary objective was to determine if a statistically significant correlation exists between the average income level of consumers and the prices of cars they purchase. The analysis employed multiple linear regression techniques to model the relationship, controlling for potential confounding variables. The results indicated a positive correlation, suggesting that higher average income levels are associated with higher car prices. The findings have implications for car manufacturers and marketers, providing insights into pricing strategies based on consumer income demographics. This study contributes to the broader understanding of economic factors influencing consumer behavior in the automotive market.

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Table of Contents

[Abstract 2](#_Toc168515313)

[1 Introduction 1](#_Toc168515314)

[1.1 Background 1](#_Toc168515315)

[1.2 Relevance 1](#_Toc168515316)

[1.3 Purpose and Questions 2](#_Toc168515317)

[2 Theory 2](#_Toc168515318)

[2.1 Regression Analysis 2](#_Toc168515319)

[2.2 Multiple Linear Regression 2](#_Toc168515320)

[2.3 Ordinary Least Squares (OLS) 2](#_Toc168515321)

[2.4 External Data Integration 3](#_Toc168515322)

[2.5 Estimation of the coefficients 3](#_Toc168515323)

[2.6 Evaluation of model accuracy 3](#_Toc168515324)

[2.6.1 Coefficient of Determination (R² and Adjusted R²) 3](#_Toc168515325)

[2.6.2 RSE and RMSE 3](#_Toc168515326)

[2.6.3 Bayesian Information Criterion (BIC) 4](#_Toc168515327)

[2.7 Stepwise Subset Selection 4](#_Toc168515328)

[2.8 Packages 4](#_Toc168515329)

[2.8.1 ggplot2 4](#_Toc168515330)

[2.8.2 MASS 4](#_Toc168515331)

[2.8.3 leaps 5](#_Toc168515332)

[2.8.4 car 5](#_Toc168515333)

[2.8.5 Metrics 5](#_Toc168515334)

[2.8.6 dplyr 5](#_Toc168515335)

[2.8.7 glmnet 5](#_Toc168515336)

[3 Method 5](#_Toc168515337)

[3.1 Data 5](#_Toc168515338)

[3.1.1 Data Collection 5](#_Toc168515339)

[3.1.2 Data Cleaning 6](#_Toc168515340)

[3.1.3 Problem and Solutions 6](#_Toc168515341)

[3.1.4 Loading Data into the R 7](#_Toc168515342)

[3.2 Modeling and Model Evaluation 7](#_Toc168515343)

[3.2.1 Modeling 7](#_Toc168515344)

[3.2.2 Questions About Group Work 9](#_Toc168515345)

[3.3 API 10](#_Toc168515346)

[4 Results and Discussions 11](#_Toc168515347)

[4.1 Model Selection 11](#_Toc168515348)

[4.2 Answers of Purpose and Questions(1.3 Section) 13](#_Toc168515349)

[5 Conclusion 14](#_Toc168515350)

[6 Theoretical questions 15](#_Toc168515351)

[7 Självutvärdering 18](#_Toc168515352)

[Searches 19](#_Toc168515353)

# Introduction

The automotive industry stands as one of the most significant sectors in the global economy, with direct impacts on both economic and social dimensions. The ever-evolving nature of the automotive market is driven by a myriad of factors, including technological advancements, consumer preferences, and economic conditions. Understanding these dynamic forces is crucial for automotive manufacturers, dealerships, and consumers alike to make informed decisions and formulate effective strategies.

## Background

The report revolves around regression analysis of the second-hand car market in Sweden. The aim is to conduct a comprehensive regression modeling based on the data collected. The data collection process involves manual extraction from blocket.se and data scraping utilizing Chrome extensions. Additionally, external data from Statistics Sweden (SCB) is retrieved via the pxweb API.

## Relevance

With the increasing global demand for automobiles, attention has been drawn to factors influencing vehicle pricing. One such factor that has become increasingly significant is the average income level within a given population or region. Theoretically, higher income levels would be expected to correlate with increased expenditures on cars and thus influence market price levels.

As an extension of the regression analysis, we will test whether average income per county affects price dynamics on the secondary market. While higher income levels generally contribute to driving prices up, we aim to empirically investigate if this holds true for used cars in the Swedish market.

En bild som visar linje, Graf, diagram, sluttning

Automatiskt genererad beskrivning

Figur 1. Cars In Sweden by Year (SCB)

## Purpose and Questions

*The purpose of this report is to investigate and analyze the relationship between average income level and car prices to shed light on how economic factors influence pricing in the automotive market.*

To fulfill this purpose, the following questions will be answered:

1. Is there a statistically significant correlation between average income level and car prices?
2. How does price sensitivity vary among consumers with different income levels within the automotive market?
3. Are there regional variations in the relationship between income level and car prices?
4. To what extent do factors other than income, such as demographic characteristics, influence car pricing dynamics?
5. Can the findings of this analysis provide actionable insights for market segmentation and pricing strategies in the automotive industry?

# Theory

## **Regression Analysis**

In the context of our research, regression analysis serves as the fundamental tool for exploring how changes in average income levels across different regions may influence vehicle prices in the second-hand car market. By examining the relationship between average income and car prices, I could quantify the impact of income on pricing dynamics, accounting for other relevant factors such as make, model, age, and mileage.

## Multiple Linear Regression

Multiple linear regression is a statistical technique used to analyze the relationship between one dependent variable and two or more independent variables. In essence, it extends the principles of simple linear regression to cases where multiple predictors influence the outcome variable.

## **Ordinary Least Squares (OLS)**

I employ Ordinary Least Squares regression to estimate the parameters of our linear regression model. This method allows us to determine the coefficients that best fit the relationship between average income and car prices, minimizing the discrepancies between observed and predicted prices while adhering to certain statistical assumptions.

## **External Data Integration**

Integrating external datasets, such as average income per county, enriches my analysis by incorporating additional variables that may influence car prices. This broader perspective allowed me to assess the interplay between income levels and pricing dynamics, providing deeper insights into the factors driving variations in vehicle prices across different regions.

## Estimation of the coefficients

In regression analysis, the coefficients represent the parameters of the model that quantify the relationship between the independent and dependent variables. Estimating these coefficients involves finding the best values that minimize the difference between the observed values of the dependent variable and the values predicted by the model. This process is typically achieved through methods such as ordinary least squares (OLS) or maximum likelihood estimation (MLE). Once the coefficients are estimated, they are used to formulate the regression equation, allowing for the prediction of the dependent variable based on the values of the independent variables. The accuracy of the coefficient estimates determines the reliability and validity of the regression model in capturing the underlying relationship between the variables.

## Evaluation of model accuracy

### **Coefficient of Determination (R² and Adjusted R²)**

The coefficient of determination (R²) measures the proportion of variance in car prices explained by variations in average income levels. Adjusted R² further refines this measure, considering the number of predictors in my model, providing a more accurate assessment of the model's explanatory power.

### RSE and RMSE

RSE (Residual Standard Error) and RMSE (Root Mean Square Error) are used to evaluate the accuracy of the model by measuring the difference between the observed values and the predicted values from the model. RSE represents the average error between the actual and predicted values, while RMSE takes the square root of RSE to indicate how much the model deviates from the actual values on average. The lower the RMSE value, the better the model fits the data.  
  
The RMSE values for both the linear regression and ridge regression models are printed to compare their performance. In this case, the linear regression model has a slightly lower RMSE compared to the ridge regression model.

### **Bayesian Information Criterion (BIC)**

I utilize the Bayesian Information Criterion to select the most appropriate model among alternatives. By balancing model fit and complexity, BIC guides my choice, ensuring that my selected model effectively captures the relationship between average income and car prices without succumbing to overfitting.

## Stepwise Subset Selection

Subset selection involves choosing the most significant predictors to improve model accuracy, reduce multicollinearity, avoid overfitting, and enhance interpretability. Stepwise selection methods offer practical alternatives to best subset selection, particularly with many variables, by exploring a limited search space.

## Packages

The packages are used while doing my regression analysis and these are all add-on packages for the R programming language that provide additional functionality and tools for performing various analyses and visualizations.

Here is a brief description of each package:

### ggplot2

A package for creating high-quality graphs and visualizations using a grammar of graphics.

### MASS

A package that contains functions for performing various statistical analyses and models, including linear regression, logistic regression, and ANOVA.

### leaps

A package used for performing variable selection in linear regression by searching through all possible combinations of predictor variables.

### car

A package that contains functions for performing various statistical analyses and models, including linear regression, ANOVA, and generalized linear models.

### Metrics

A package that contains functions for calculating various statistical metrics, including RMSE, MAE, and R-squared.

### dplyr

A package used for manipulating and filtering dataframes in an efficient manner. *(R programmering packages)*

### glmnet

A package used for performing regularized linear regressions and classifications using Lasso and Ridge regression techniques.

En bild som visar text, Teckensnitt, skärmbild, design

Automatiskt genererad beskrivning

Figur 2. Installing packages

# Method

## Data

### Data Collection

When we started the project, we collected the data manually from Blocket.se according to the region preferred by each one in my group. The collected cars were filtered as 'electric cars' and chose accordingly and put on Excel document by each of group member and then combine to one Excel document. The collected and combined data were seen in Excel document as the following:

En bild som visar text, skärmbild, nummer

Automatiskt genererad beskrivning

Figur 3. The collected and combined data in Excel

### Data Cleaning

When data collection, data cleaning was performed due to the differences in the collected data. Before this process, there were 214 observations. We noticed that some of the values are not the same as other values. This is categorically causing error so we need to clean the data that we collected. After the preprocessed in the same value after we gathered again in the excel document. We started my regressionanalysis with cleaned data. We had totally 1047 observations. The collected, cleaned and preprocessed data consisted of these coloums:

|  |  |  |
| --- | --- | --- |
| * Location * Year * Name * Miles | * Gear * Model * Brand * Fuel | * Engine Volume * Horsepower * Price |

### Problem and Solutions

In order to use the data collected in Excel efficiently in R, we examined the data in detail and edited the missing or incorrectly entered data. Here are the errors we found in the data that required cleaning:

* **Price:** If the price was not entered and there were spaces between the numbers, we closed the spaces and entered it in the same way.
* **Name:** Some vehicles with the same name were entered incompletely or incorrectly by the person who entered the announcement. We found similar names and entered them with the same name to prevent any differences.
* **Miles:** Some of the vehicles with 0 miles did not have data, we completed them by entering 0. These were especially present in 2024 model vehicles.

### Loading Data into the R

When the data cleaning was done, the dataset is loaded, and a summary of its contents is displayed. Data is filtered to exclude entries with prices below 30,000 swedish krona and rows with missing values are removed to ensure data quality.

**By Exploratory Data Analysis (EDA) created** histograms of the variables Price, Horsepower, Year, and Miles using the hist function to visualize the distribution of the data which helps in identifying potential transformations (e.g., taking the log of mileage).

En bild som visar text, diagram, skärmbild, linje

Automatiskt genererad beskrivning

Histograms of key variables (”Price”, ”Horsepower”, ”Year”, ”Miles”, and log-transformed ”Miles”) are plotted to visualize their distributions. A frequency table of the ”Location” variable is printed to understand the distribution of car listings across different locations.

## Modeling and Model Evaluation

### Modeling

The code creates a linear regression model using the lm function with the formula "Price ~ . -Brand -Model -Engine.Volume -Name -Company" to predict the price of a car based on the other variables in the dataset. The model is assigned to the variable "model\_l". I split the dataset into training and test sets using the sample function.

En bild som visar text, diagram, skärmbild, linje

Automatiskt genererad beskrivning

Figur 4. Linear Regression Model Diagnostics Plots

Diagnostic plots of the linear regression model using the plot function and sets the layout to 2x2 using the par function.

After modeling I created a new variable "Age" by subtracting the ”Year” variable from 2024 and creates four new variables by multiplying Miles and Age, dividing Miles by Age, multiplying Horsepower and Age, and taking the logarithm of Age.

En bild som visar text, skärmbild, diagram

Automatiskt genererad beskrivning

Figur 5. Age Feature Plots

First plot shows the relationship between the age of the car and its price. Typically, we expect to see a negative correlation here, as older cars tend to be less expensive due to depreciation. If the plot shows a downward trend, it confirms this expectation. Second plot examines the interaction between the car's age and its mileage on the price. The combined effect of a car being older and having higher mileage is usually more significant on the price. A downward trend would indicate that cars which are both older and have higher mileage are valued lower. Third plot shows transforming the age variable if a logarithmic relationship better describes the data. A downward trend in this plot would still suggest that older cars (on a logarithmic scale) tend to be less expensive. And the last plot shows the relationship between the price and the ratio of mileage to age, effectively the average miles per year. A downward trend would indicate that cars more extensively on a yearly basis are valued lower, which can be an indicator of more wear and tear.

After buildning linear regression model I build several linear regression models to predict car prices. A basic linear model without interactions. A log-transformed linear model to address skewness in price data. External data on average income by region is added to the dataset. An extended log-transformed linear model includes average income by region as a predictor.

Here is all linear regression models, each with slight variations and different specifications:

1. **model\_l**: This is a basic linear regression model that includes all predictors except for "Brand," "Model," "Engine.Volume," "Name," and "Company."
2. **model\_l\_int**: Similar to the first model, but it includes interaction terms between "Horsepower" and "Age" and between "Miles" and "Age," while excluding the same variables as the first model.
3. **model\_l\_log**: This model differs from the previous ones as it predicts the natural logarithm of the price (log(Price)), which is achieved by transforming the dependent variable. It excludes "X" (assuming it's not a predictor variable) and includes interaction terms between "Miles" and "Age."
4. **model\_l\_log\_ext**: Similar to the third model, but it also excludes "Location" and includes the interaction term between "Miles" and "Age." Additionally, it appears to use a modified dataset named "bil\_data\_clean\_1."

### Questions About Group Work

1. Who have you worked in a group with?

Dan Heikenberg, Daniel Hemgren, Frida Kilby, Khaldoun Agha, Natalie Dobrovoska, Sierhei Thor Fedatsenka, William Blennow, Xiaoyong Yang.

1. How have you in the group worked together?

We collected the data manually from Blocket.se according to the region preferred by each one in my group. The collected cars were filtered as 'electric cars' and chose accordingly and put on Excel document by each of group member and then combine to one Excel document.

1. What was good in the group work and what can be developed?

We helped each other, for example; we worked together to identify the differences we noticed in the data and what to do to resolve them. We also made joint decisions about what to do, thanks

to good communication. And I think this is an important thing for group work. Although decisions for greater interoperability as a group were common, individual work was at the forefront.

1. What are your strengths and development opportunities when you work in a group?

I find myself careful and successful in spotting mistakes and problems. When I manage to solve problems, I aim to do better together by sharing this information and solutions.

1. Is there anything you would have done differently? What in such cases?

While collecting data, I realized that I entered the records without errors and with the same value.

## API

The data that was decided to be retrieved from SCB was total cars in Sweden per year. To retrieve external data, I used the statistics database API (PXWEBAPI 1.0). By processesing in Python programmering with using JSON command and URL was smooth and effective way to retrieve data. Necessary modules for plotting and custom tick formatting. Example; Converts y-axis values to integers(total cars) with comma separators in plot(introduction section). This ensures that when the data is plotted, the y-axis values are displayed enhancing the clarity of the visualization. The data before plotting is in the below:

En bild som visar text, Teckensnitt, skärmbild, typografi

Automatiskt genererad beskrivning

Figur 6. Data

# Results and Discussions

## Model Selection

The models aim to predict car prices based on various features, potentially improving accuracy by incorporating regional income data. Before selecting the best model, first thing to know that lower RSE values indicate a better fit. Adjusts R-squared for the number of predictors; useful for comparing models with different numbers of predictors.

|  |  |  |  |
| --- | --- | --- | --- |
| model\_l | model\_l\_int | model\_l\_log | model\_l\_log\_ext |
| 0.7676 | 0.8152 | 0.8561 | 0.8544 |

Tabell 1. Models' R^2 Errors

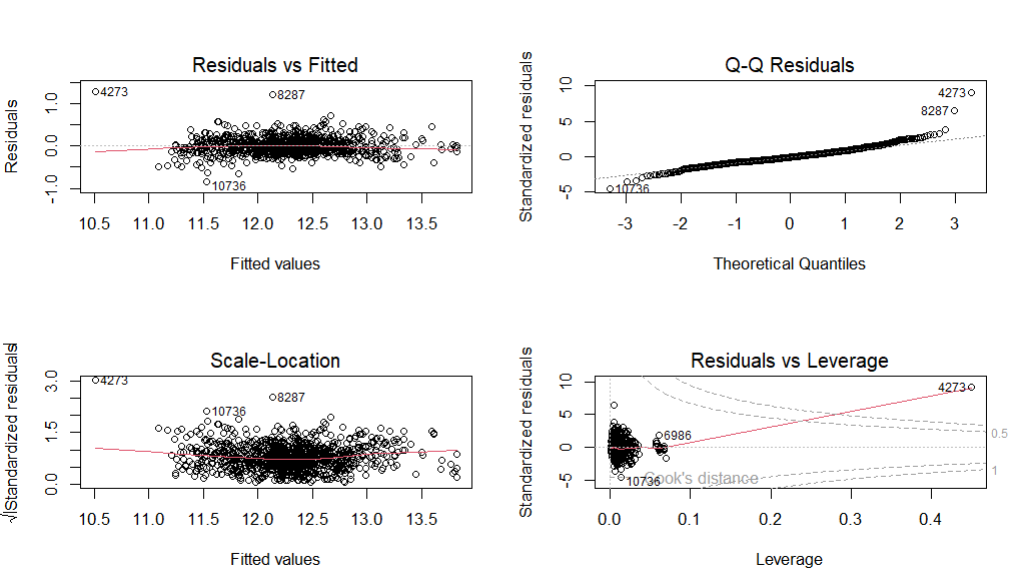
After comparing the models according to R^2 model\_l\_log is the best performing model in terms of fit (R-squared values) and has a very low residual standard error on the log-transformed price. However, model\_l\_log\_ext is simpler with fewer predictors while maintaining almost the same level of performance as model\_l\_log. And model\_l\_log\_ext has the highest R^2 (0.8544), indicating the greatest proportion of variance explained by the independent variables. So model\_l\_log\_ext is the best model can be selected.

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Figur 7. Metrics of model\_l\_log\_ext

Based on these metrics, model\_l\_log\_ext appears to be the best model as it has lowest RSE (0.187), indicating the smallest average deviation of observed values from the fitted values. And the model also has highest R^2 and highest F-statistic (683.2) and a p-value < 2.2e-16, indicating high overall model significance.



Figur 8. Diagnostic Plot of Selected Model(model\_l\_log\_ext)

* Significant predictors include miles, gear, horsepower, car age, and the interaction between miles and age.
* The average income level of different regions is not a significant predictor of car prices.

According to the model, the coefficient for Mean\_income is -1.754e-04, and it is not statistically significant (p = 0.19468).

This indicates that there is no strong evidence of a relationship between car prices and the average income level in different regions based on this dataset.

## Answers of Purpose and Questions(1.3 Section)

These questions were obtained and answered by regression analysis and this report.

1. **Is there a statistically significant correlation between average income level and car prices?**

In my regression analysis, I found that the average income level across different regions did not show statistical significance in determining car prices. This suggests that there may not be a strong correlation between income level and car prices in my dataset.

1. **How does price sensitivity vary among consumers with different income levels within the automotive market?**

My analysis did not directly address price sensitivity among consumers with different income levels. However, it's possible to infer that if income level doesn't significantly impact car prices, the price sensitivity across income groups might not vary significantly within the automotive market.

1. **Are there regional variations in the relationship between income level and car prices?**

My analysis primarily focused on national-level data and did not delve into regional variations. However, future research could explore regional differences to determine if there are variations in the relationship between income level and car prices across different areas.

1. **To what extent do factors other than income, such as demographic characteristics, influence car pricing dynamics?**

My regression analysis considered various factors such as age, mileage, horsepower, and fuel type, alongside income level. These factors showed varying degrees of influence on car prices. Demographic characteristics like age and mileage appeared to have a stronger predictive ability on car prices compared to income level in my models.

1. **Can the findings of this analysis provide actionable insights for market segmentation and pricing strategies in the automotive industry?**

While income level may not directly impact car prices in my analysis, understanding the influence of demographic factors, mileage, and age on pricing dynamics can provide valuable insights for market segmentation and pricing strategies. For instance, targeting specific consumer segments based on their preferences for mileage, age of the vehicle, or fuel type could help optimize pricing strategies and enhance competitiveness in the automotive market.

# Conclusion

The final regression model demonstrates strong predictive power and provides valuable insights into the factors determining car prices. The significant predictors include fuel, mileage, horsepower, age, and their interactions. The analysis highlights the importance of considering both individual and combined effects of predictors on car prices.

According to the regression analysis of the dataset, it appears that the average income per district does not have a significant effect on automobile prices.

The regression analysis reveals that car prices are significantly influenced by factors like fuel, miles, gear, horsepower, and age. The average income level in different regions does not have a significant impact on car prices in this dataset. The interaction between miles and age suggests that older cars depreciate more slowly with additional miles compared to newer cars.

Overall, the findings from this analysis can aid stakeholders in the automotive industry in understanding the key drivers of car prices, enabling more informed pricing strategies and market analysis.

For further research could focus on collecting more data, particularly for electric cars and different income levels, to enhance the model's robustness and predictive capability. Additionally, exploring non-linear relationships and employing advanced regression techniques like Ridge or Lasso regression could further improve the model’s performance.

# Theoretical questions

1. Check out the following video: https://www.youtube.com/watch?v=X9\_ISJ0YpGw&t=290s , briefly describe what a Quantile-Quantile (QQ) plot is.

A Quantile-Quantile (QQ) plot is a graphical method for comparing the distribution of two datasets. In a QQ plot, the quantiles of one dataset are plotted against the quantiles of the other dataset. If the two datasets have the same distribution, the points in the QQ plot will lie along a straight line. If one dataset deviates from the other, the points will deviate from the line. The QQ plot is often used to investigate whether a dataset has a certain theoretical distribution, such as a normal distribution, by comparing its quantiles with the quantiles of the theoretical distribution.

1. Your colleague Karin asks you the following: "I have heard that in Machine Learning the focus is on predictions, while in statistical regression analysis you can make both predictions and statistical inference. What does that mean, can you give some examples?” What do you answer Karin?

In machine learning, the focus is on building models that can make predictions with high accuracy on new data. This means that techniques such as cross-validation are often used to evaluate the models' prediction ability. In statistical regression analysis, it is also possible to make predictions, but it is also common to want to draw conclusions about how different factors affect the outcome variable. This means that statistical inference techniques, such as hypothesis testing or confidence intervals, are often used to assess whether a factor has a significant effect on the outcome variable or not. An example could be investigating whether a certain medication has a significant effect on a disease, and then statistical inference techniques can be used to assess whether the effect is real or just a coincidence.

1. What is the difference between "confidence interval" and "prediction interval" for predicted values?

Confidence interval and prediction interval are two different types of intervals used to assess the uncertainty in predictions of an outcome variable based on a model. The confidence interval is an interval that contains the true mean of the outcome variable with a certain probability, while the prediction interval is an interval that contains the expected value of a single observation of the outcome variable with a certain probability. The confidence interval describes the uncertainty in the mean of the outcome variable, while the prediction interval describes the uncertainty in a single observation of the outcome variable.

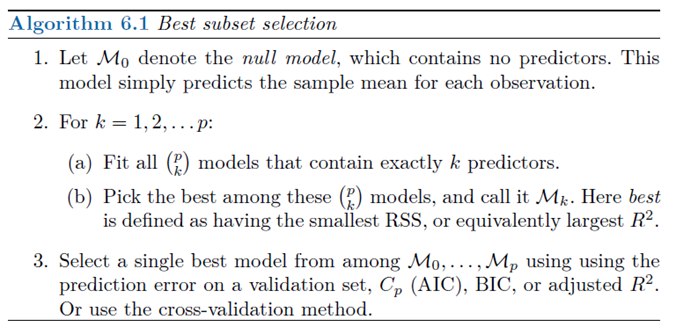
1. The multiple linear regression model can be written as: 𝑌 = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ . . . + 𝛽𝑝𝑥𝑝 + 𝜀 How are the beta parameters interpreted?

In the multiple linear regression model the 𝛽 parameters are regression coefficients that describe the expected change in the outcome variable 𝑌 for each unit change in the respective explanatory variable 𝑥. More specifically, the 𝛽 parameters are interpreted as follows: 𝛽0 is the intercept, the expected value of 𝑌 when all explanatory variables are zero. 𝛽1 is the regression coefficient for the explanatory variable 𝑥1, the expected change in 𝑌 for each unit change in 𝑥1, while all other explanatory variables are held constant. 𝛽2 is the regression coefficient for the explanatory variable 𝑥2, the expected change in 𝑌 for each unit change in 𝑥2, while all other explanatory variables are held constant.

1. Your colleague Hassan asks you the following: "Is it true that in statistical regression modeling you do not need to use training, validation and test sets if you use measures such as BIC?" What is the logic behind this?” What do you answer Hassan?

Yes, it is true that measures such as BIC can be used as a criterion for model selection in statistical regression modeling. The logic behind BIC and similar methods is that they penalize excessive complexity and favor simpler models that still perform well. While using test sets is not necessarily required when using BIC, it is still recommended to evaluate the model's performance on new data. In summary, while BIC can be used to select the best model among a set of candidate models, it is still important to evaluate the model's performance on new data using test sets.

1. Explain the algorithm below for "Best subset selection"



Best subset selection is a method for selecting the best set of predictors for a linear regression model by testing all possible combinations of predictors. The algorithm starts with a null model and then considers all possible models that can be formed by adding one predictor variable at a time. The best model is selected based on a criterion such as the adjusted R-squared or the BIC. This process is repeated for all possible combinations of predictor variables up to a specified maximum number of variables or until a stopping criterion is met. Best subset selection always gives the best combination of predictors but can be computationally intensive, especially with a large number of predictors. Therefore, it is not practical for large datasets.

1. A quote from statistician George Box is: “All models are wrong, some are useful.” Explain what is meant by that quote.

The quote "All models are wrong, some are useful" by George Box means that models are not completely accurate and are simplified representations of reality. However, models can still be useful for understanding and predicting real phenomena, identifying important factors, testing hypotheses, and drawing conclusions about relationships between variables. It is important to be aware of the model's limitations and to validate the model on unseen data to assess its prediction ability on new data.

# Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Det var en av de svåraste kurserna för mig att fokusera på eftersom jag blev sjuk och inte kunde fokusera på ämnena, så jag kunde inte avsluta de två delarna av kunskapskontroll tillsammans med kursen, och medan jag fortsatte med Deep Learning-kursen, det var svårt att försöka komplettera mina brister, lära sig, förbättra och förbereda rapporten, men nu känner jag mig lättad över att jag lyckades. Jag vill också tacka Antonio för att han var flexibel och hjälpte mig i detta avseende på grund av min situation.

1. Vilket betyg du anser att du skall ha och varför.

Jag gjorde också API-delen och med allt jag gjorde var VG-kriterierna uppfyllda, naturligtvis, det slutliga utrymmet tillhör Antonio. Han lät mig skicka in det sent. Jag hoppas att detta inte påverkar mina betyg för trots allt kämpade jag mycket.

1. Något du vill lyfta fram till Antonio?

Det är inget jag vill nämna specifikt. Tack för allt 😊

# Searches

* Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (Second Edition)*. Sebastopol, CA: O’Reilly Media, Inc.
* James G. , Witten D. , Hastie T. ,Tibshirani R. (2023). An introduction to statistical learning : with applications in R. New York, Springer
* Statistikdatabasens API - [Statistikdatabasens API (PxWebApi 1.0) (scb.se)](https://www.scb.se/vara-tjanster/oppna-data/api-for-statistikdatabasen/)
* R programmering packages - [Packages in R Programming - GeeksforGeeks](https://www.geeksforgeeks.org/packages-in-r-programming/?ref=lbp)