R-programming

Building a secondhand car price predicting model

Shriya Walia

EC Utbildning

Kunskapskontroll- R

2024-04

Abstract

The objective of this thesis is to construct a robust secondhand car price prediction model utilizing linear models in R. The methodology encompasses several key stages, including data cleaning, feature engineering, model evaluation, and model selection. Through a systematic workflow, the methodological approach, results, and conclusions have been thoroughly deliberated, resulting in the identification of the Lasso Regression model as the most suitable candidate. The thesis delves into each phase of the process, detailing the techniques employed and the rationale behind the decisions made, ultimately providing a comprehensive framework for developing an effective predictive model for secondhand car prices.

Content

Outline of the Thesis

The thesis commences with an introduction outlining the project's scope, followed by explanation of theoretical concepts essential for understanding the undertaken project. Subsequently, the methods employed, and the results obtained are discussed. A conclusive summary is provided, highlighting the insights received from the project, suggestions for enhancing performance, and a discussion of the challenges encountered.

1 Introduction

Regression analysis traces its roots to the late 19th century when Sir Francis Galton introduced the concept of regression to the mean while examining hereditary traits. Building upon Galton's groundwork, Karl Pearson and Sir Francis Ysidro Edgeworth played pivotal roles in shaping modern regression analysis in the early 20th century.

Regression, often unnoticed, permeates our daily lives, shaping outcomes in many areas. Its influence extends far beyond statistical analyses, from health, finance, meteorology, and beyond. The world operates within the framework of statistical principles, with regression serving as an unseen force shaping outcomes and restoring equilibrium across diverse systems and disciplines.

The increasing trend in car ownership, as depicted in Figure 1, made by car data collected from Statistikmyndigheten SCB using API, underscores the growing significance of understanding and effectively pricing second-hand cars. With more cars being purchased annually, the market for used vehicles naturally expands, presenting both buyers and sellers with the challenge of accurately determining fair prices. This task is complicated by numerous factors such as depreciation, mileage, age, model year, brand reputation, and market demand. In such a complex landscape, the need for sophisticated modeling techniques becomes apparent. **Multiple linear regression emerges as a powerful solution, enabling the analysis of relationships between multiple independent variables—such as mileage, age, horsepower, and more—and the dependent variable of car price.** By developing a multiple linear regression model, valuable insights into pricing trends can be gained, empowering both buyers and sellers to make informed decisions in a dynamic and expanding market. Through predictive modeling, stakeholders can better anticipate fluctuations in car prices, identify factors driving these changes, and adjust their strategies accordingly, ultimately fostering a more efficient and transparent marketplace for second-hand cars.

This thesis aims on training and selecting a multiple linear regression model for predicting prices of secondhand cars while answering the following questions:

1. What challenges arise while preprocessing data?
2. Which linear model works best on secondhand car data?
3. What are the factors influencing the model?
4. What does feature selection help with?

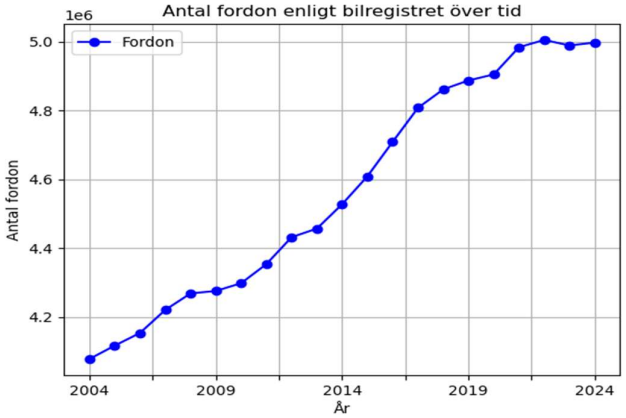


Figure 1: Increase in number of personal cars in Sweden from 2004-2024

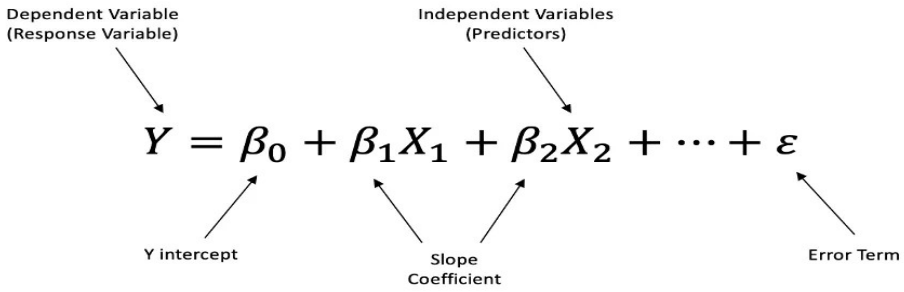
2 Theory

Here we will go through the concepts used in this thesis:

2.1 Multiple Linear Regression

Multiple linear regression, as described by Taylor (n.d.), refers to a statistical technique used to predict the outcome of a variable based on the value of two or more variables. It is an extension of linear regression and is sometimes known simply as multiple regression. In this method, the variable to be predicted is termed the dependent variable, while the variables used to predict its value are referred to as independent or explanatory variables.

The multiple linear regression formula:



The multiple linear regression model relies on several key assumptions:

1. **Linear Relationship**: It assumes a linear relationship between the dependent variable and the independent variables.
2. **Low Multicollinearity**: The independent variables should not be highly correlated with each other to avoid multicollinearity issues.
3. **Constant Variance of Residuals**: The variance of the residuals should remain constant across all levels of the independent variables, ensuring homoscedasticity.
4. **Independence of Observations**: Each observation in the dataset should be independent of the others, with no systematic patterns or dependencies among them.
5. **Normality of Residuals**: The residuals, which represent the difference between the observed and predicted values, are assumed to follow a normal distribution.

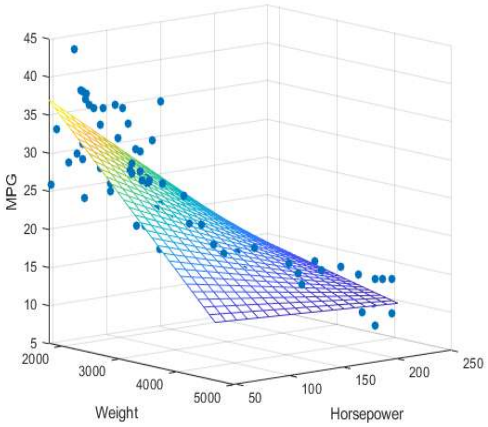


Figure 2: Multiple Linear Regression graph

2.2 Lasso Regression

Lasso regression is a regularization technique that applies a penalty to prevent overfitting and enhance the accuracy of statistical models IBM. (n.d.). In multiple linear regression, overfitting can occur when the number of predictors is large relative to the number of observations or when predictors are highly correlated. Lasso regression addresses this issue by introducing a penalty term that is proportional to the absolute values of the regression coefficients. This penalty encourages simpler models by shrinking some coefficients to zero, effectively performing variable selection and regularization simultaneously. As a result, Lasso regression selects a subset of the most important predictors while setting others to zero, leading to a more interpretable model with improved prediction accuracy and generalization performance. **By striking a balance between bias and variance, Lasso regression provides a powerful tool for feature selection and regularization in multiple linear regression, helping to mitigate the effects of multicollinearity and enhance the robustness and interpretability of the regression model.**

Lasso Regression formula:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$$

Greater lambda (**λ**) leads to greater penalty as shown in Figure 3.

The first part of the objective function represents the sum of squared differences between the observed and predicted values of the dependent variable. The second part of the objective function represents the penalty term, which penalizes the absolute values of the regression coefficients. The penalty term is multiplied by the regularization parameter λ, which controls the degree of regularization applied to the model.

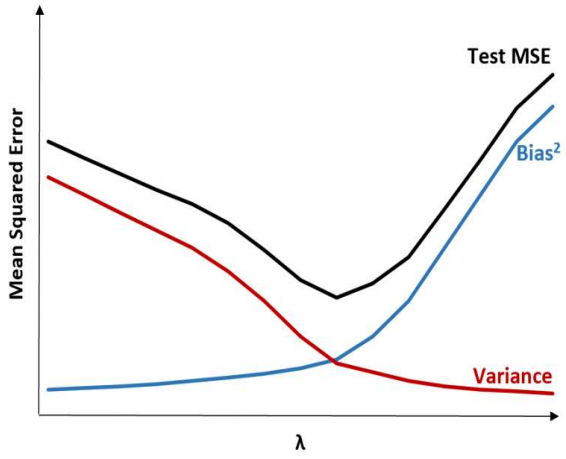


Figure 3: Working of lambda (λ) in Lasso Regression

2.3 Evaluation Metrics

2.3.1 R-squared (R^2)

R-squared measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. Higher values of R-squared indicate a better fit of the model to the data. R squared is usually not a good metric when comparing multiple linear regression models since the more the variables added, the r squared will increase. Therefore, the concept of adjusted R-squared is used more.

2.3.2 Residual standard error or RMSE (Root Mean Squared Error)

RMSE measures the average difference between the observed and predicted values of the dependent variable. Lower RMSE values indicate better predictive accuracy.

2.3.3 Adjusted R^2

Adjusted R-squared is similar to R-squared but penalizes for the number of predictors in the model, making it more suitable for comparing models with different numbers of predictors like comparing multiple linear regression models with different number of features.

2.3.4 BIC (Bayesian Information Criterion)

BIC serves as a model selection criterion, striking a balance between model fit and complexity by penalizing models with a large number of parameters. A lower BIC value indicates a superior equilibrium between model fit and complexity.

$$BIC = n \cdot \ln(\sigma^2) + k \cdot \ln(n)$$

Where:

- n is the number of observations.
- σ^2 is the estimated variance of the residuals.
- k is the number of estimated parameters in the model.

2.3.5 MAPE (Mean Absolute Percentage Error)

MAPE is a measure of the accuracy of a forecast or prediction model. It calculates the average absolute percentage difference between predicted values and actual values.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100$$

Where:

$\hat{y}_i$ = Predicted value for the i^{th} data point

y_i = Actual value for the i^{th} data point

n = number of observations

3 Method

The process of selecting the best multiple regression model entails data collection, outlier detection and removal, feature selection, model evaluation and selection for the secondhand car price predicting models.

3.1 Tools

The coding was executed in **R Studio**, enabling the utilization of essential statistical tools such as evaluation metrics (RMSE, adjusted R-squared, etc.) on the linear models to assess and compare their performance. For the API integration, Python was employed within Jupyter Notebooks to extract external data.

3.2 Data

The data for the project on secondhand cars was manually collected from 'Blocket', a platform for buying and selling pre-owned items, by a team of Data Science students. The data collection process involved assigning different cities in Sweden to each team member, with each member responsible for gathering at least 100 data points. This collaborative effort resulted in a dataset comprising 700 data points, including details such as price, city, fuel type, gearbox, mileage, model year, car type, driving type, horsepower, color, and brand. Working as a group proved beneficial, as discussions allowed for the exchange of ideas and suggestions, enhancing the overall quality of data collection. An observed personal growth aspect within the group dynamics was the heightened motivation and improved outcomes compared to working individually.

Furthermore, external data integrated into the project was sourced from SCB (Statistics Sweden). This supplementary dataset provided information on the number of personal cars registered in Sweden annually, specifically in the month of March, spanning from 2004 to 2024.

3.3 Processing of data

Firstly, the data underwent cleaning, which involved removing unnecessary columns such as county name, motor size, and date in traffic. Subsequently, each column's contents were scrutinized for repetitions and corrected as needed. Once the data was cleaned in Excel, it was imported into R for further preprocessing. To facilitate analysis, the categorical columns were converted into dummy variables using the 'fastDummies' library, as dealing with numerous categorical variables can pose challenges, especially when using functions like VIF (Variance Inflation Factor) or Lasso regression models. Diagnostic plots were then generated using linear models to assess the data distribution, as depicted in Figure 4. The diagnostic plots revealed numerous outliers, prompting their removal. Subsequent iterations of diagnostic plot checks and outlier removal were conducted until no outliers remained, and all plots exhibited satisfactory characteristics. Specifically, the residuals vs. fitted plot showed constant and uniform variance, the QQ plot exhibited a normally distributed pattern with residuals aligning closely with the curve, and the residuals vs. leverage plot did not indicate any data points with

a Cook's distance exceeding 1 or exerting undue influence. Upon confirming these conditions, as illustrated in Figure 5, further steps were undertaken.

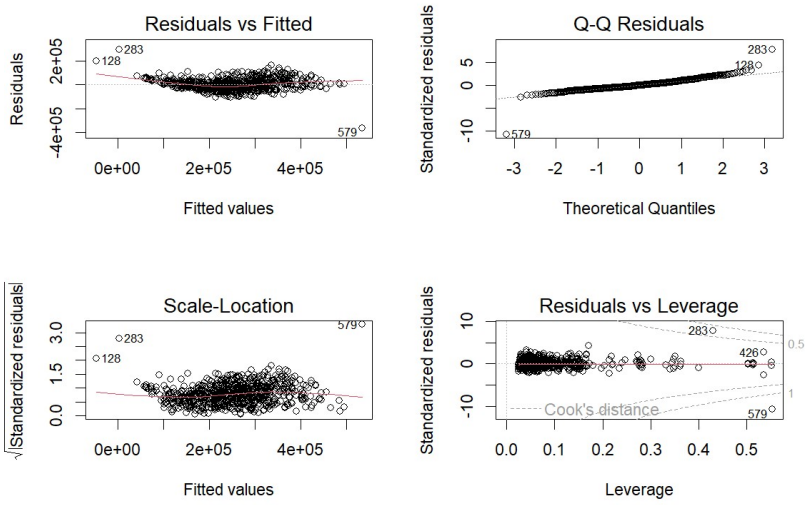


Figure 4: Diagnostic plots before data cleaning

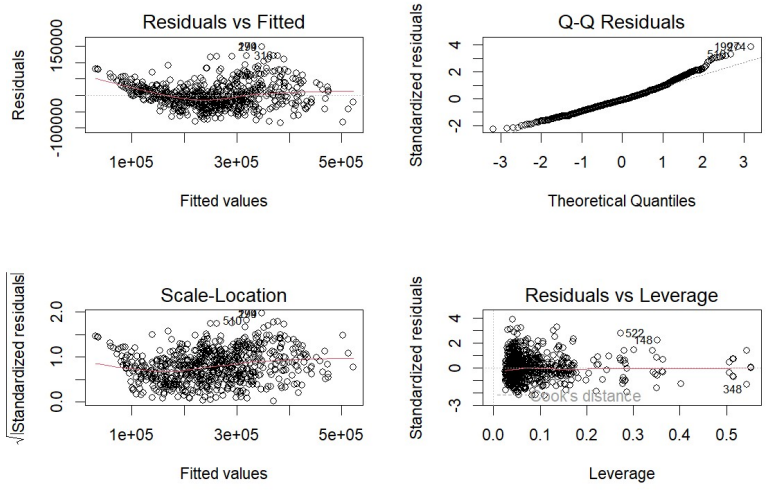


Figure 5: Diagnostic plots after data cleaning

3.4 Splitting of data

The data was divided into three parts: the training set was used to compare the models, the validation set was employed to assess how the models performed on unseen data, and the test set was utilized to evaluate the performance of the best models.

3.5 Feature selection

The features were selected with the help of the training data using the lm() function to fit linear regression model by the following steps:

- After initially including all features in the model 'lm_train', I observed a high correlation between 'mileage' and 'model year', prompting the removal of 'mileage' from the dataset.
- However, subsequent examination using the VIF (Variance Inflation Factor: a measure of the amount of multicollinearity in regression analysis) on 'lm_train_1_1' revealed high VIF values (>10) for 'car types' and some 'brands', indicating multicollinearity issues. As a result, I removed the 'car type' features from the data and retrained the model, yet multicollinearity persisted.
- In another attempt, I retained the 'car type' and removed brands with high VIFs. However, high VIF values persisted even for the car type features. Consequently, I opted to remove both the 'car type' and 'brands' with high VIFs, leading to a dataset without any features exhibiting high multicollinearity.
- Subsequently, I conducted best subset selection using the 'leaps' library on this refined dataset, ultimately selecting 5 predictors that resulted in an optimal adjusted R-squared, as depicted in Figure 6 on the model 'lm_train_with5'.

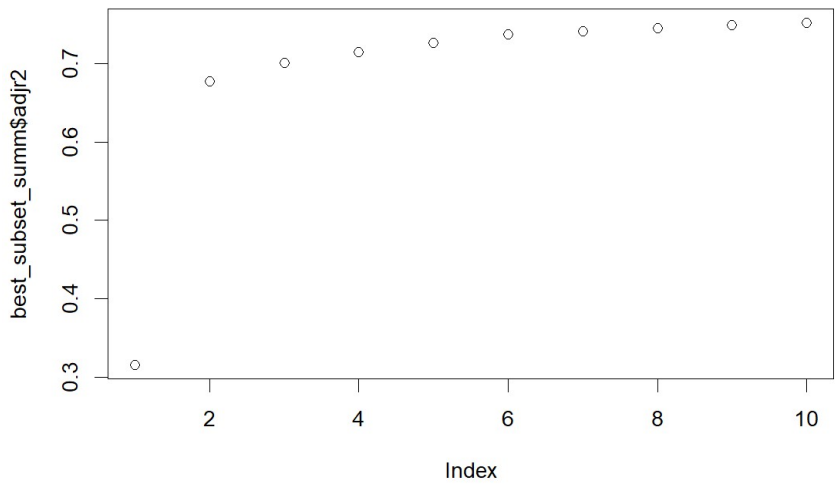


Figure 6: Best subset selection

3.6 Model evaluation

Model evaluation was conducted on the 6 models selected from the previous feature selection experiment, on the validation data & test data. These models were compared based on their performance metrics, including RMSE (Root Mean Squared Error), adjusted R-squared, MAPE (Mean Absolute Percentage Error) and BIC (Bayesian Information Criterion) using the 'Metrics' library. The purpose of this evaluation was to assess how well each model performs in terms

of predictive accuracy and goodness of fit. By comparing these metrics across the different models, we can determine which model provides the best balance between model complexity and predictive performance.

3.7 Lasso Regression

Furthermore, in continuation of the model evaluation process, a Lasso regression model was trained using the 'glmnet' library on the training dataset containing all features, without employing any feature selection techniques. The primary purpose was to utilize Lasso regularization, which automatically performs feature selection by shrinking some coefficients to zero based on the optimal lambda value. Subsequently, the performance of the Lasso model was assessed on both the validation and test datasets using RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error) metrics. These metrics were calculated using the Metrics library.

4 Result

The results of the best secondhand car price predicting model are as follows:

4.1 Selecting the best perfoming model

The best-performing model was selected by evaluating their performance on both the validation and test sets using RMSE and MAPE. Adjusted R-squared and BIC metrics were utilized on the training data. Here are the results:

Model	RMSE	MAPE	Adjusted R-squared	BIC
Model 1 (all features) **Validation**	43000.89	13.94 %	84.2%	1011.23
Test	44712.56	15.74%		
Model 2 (removed mileage) **Validation**	46737.86	15.65%	80.89%	10185.44
Test	50404.57	18.08%		
Model 3 (without mileage + car types with high VIF) **Validation**	50889.79	16.78%	77.53%	10232.20
Test	54346.76	18.95%		
Model 4 (without mileage + car brands with high VIF) **Validation**	47433.67	15.79%	79.52%	10179.79
Test	50021.23	18.20%		
Model 5 (without mileage + car brands+ car brands & types with high VIF) **Validation**	52068.34	17.47%	75.74%	10234.33
Test	53998.47	19.31%		
Model 6 (with 5 features selected by best subset selection) **Validation**	52848.04	16.93%	72.6%	10062.03
Test	51245.20	18.36%		
Lasso Regression Model **Validation**	42044.48	13.64%	NA	NA
Test	42802.33	14.46%	NA	NA

Based on these results, several observations can be made:

- Model 1 (All Features) performs reasonably well, with relatively low RMSE and MAPE on both validation and test sets. However, it seems to suffer from overfitting as indicated by the relatively high Adjusted R-squared since feature selection has not been done.
- The Lasso Regression Model appears to perform well, with the lowest RMSE and MAPE on the validation set.
- Models 3, 4, and 5, which involve removing features with high VIF, show mixed performance compared to Model 1.
- Model 6, with feature selection via Best Subset Selection, shows relatively higher RMSE and MAPE compared to other models.
- Considering the trade-off between model complexity (as indicated by BIC) and predictive performance, **the Lasso Regression Model is most preferable choice due to its simplicity and good performance on the validation set & test set**. Close competitor to Lasso Regression Model would be Model 4.

4.2 Model Performance example on real life data

The best model i.e., the Lasso Regression Model was used to predict the price of a random secondhand car picked from Blocket, and the results are as follows:

Figure 7: Secondhand car on Blocket

The Lasso Regression model has seemed to predict the price of the second hand car well, with just a difference of approx, 12,000 kr more than the price on blocket, set by the seller.

Price predicted by Lasso Regression model :

```
Predicted Price of the Car: 251285.3
```

5 Conclusion

The conclusion of this thesis answers the questions asked in the beginning:

5.1 Challenges faced while preprocessing data & model limitations

The data collection process was constrained by time limitations, resulting in data being gathered from a select few cities in Sweden with only 100 data points per location. This restricted availability of data due to limited resources. After initial cleaning in Excel, diagnostic plots revealed numerous outliers, primarily stemming from cars with either very old model years or exceptionally high horsepower. Further constraints included limiting the dataset to a specific price range of 100k to 500k, excluding family cars and work vehicles. Consequently, the models created were designed to accommodate these specific features and limitations present in the dataset.

5.2 Factors influencing the model

A model's performance can be influenced by several factors. Correlated features and multicollinearity can undermine the model's accuracy. Additionally, problematic diagnostic plots, such as high leverage points, outliers, and non-normally distributed residuals, can contribute to overfitting and ultimately result in poor performance. Addressing these issues is crucial before drawing conclusions about the best model.

5.3 Feature selection

Feature selection plays a vital role in evaluating how various combinations of features impact model performance. Through feature selection, the process of identifying the most influential features becomes more robust and reliable as different feature combinations are explored. Despite undergoing feature selection in this project, the Lasso regression model ultimately outperformed others by autonomously selecting the most significant features.

5.4 API

The external data from the SCB platform was collected using Python code in Jupyter Notebooks, allowing the extraction of necessary information. This included generating the graph (Figure 1), which illustrates the trend of personal car ownership in Sweden over the past decade. This visualization adds significance to the project by highlighting the growing demand for a reliable secondhand car price prediction model.

The selected model shows considerable potential but requires extensive training to further enhance its performance. With more training, it can achieve even better results. Alternatively, if considering other models, they could undergo additional training to become better suited for real-world applications.

6 Teoretiska frågor

1. Beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

Svar: Inom statistik är en Q–Q-plot (kvantil–kvantilplot) en sannolikhetsplot som kontrollerar om en given datauppsättning följer en viss sannolikhetsfördelning, typiskt normalfördelningen. Det är en grafisk metod för att jämföra två sannolikhetsfördelningar genom att plotta deras kvantiler mot varandra. Om punkterna på diagrammet faller ungefär längs en rät linje, tyder det på att datasetet är väl modellerat av den valda fördelningen.

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?

Svar: Ja, det är korrekt att Machine Learning sysslar med att bygga prediktiva modeller som hjälper till att prognostisera utfall baserat på indata. Medan i statistisk regressionsanalys görs statistiska slutsatser om prediktioner och gör det möjligt för oss att dra slutsatser om betydelsen av individuella prediktorvariabler samtidigt som vi bedömer modellens övergripande passform. Till exempel kan en maskininlärningsalgoritm förutsäga sannolikheten för att en patient har en viss sjukdom baserat på sina symtom, medan regressionsanalys kan användas för att utforska sambandet mellan olika riskfaktorer (såsom ålder, livsstilsvanor, genetik) och förekomsten av en speciell sjukdom.

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

Svar: Ett konfidensintervall för ett predikterade värde representerar det intervall inom vilket vi förväntar oss att det sanna medelsvaret (eller förväntade värdet) ska falla med en viss konfidensnivå. Om vi till exempel predikterar den genomsnittliga lönen för en stor befolkning, till exempel 25-åringar, skulle ett konfidensintervall ge ett intervall inom vilket vi är säkra på att den verkliga genomsnittslönen för alla 25-åringar i befolkningen ligger.

Å andra sidan representerar ett prediktionsintervall för ett predikterade värde, det intervall inom vilket vi förväntar oss att en individuell framtida observation faller med en viss konfidensnivå. Om vi till exempel predikterar lönen för en specifik 25-årig individ, skulle ett prediktionsintervall ge ett intervall inom vilket vi är säkra på att deras faktiska lön kommer att falla.

4. Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$. Hur tolkas beta parametrarna?

Svar: Varje β-parameter i den multipellinjära regressionsmodellen representerar förändringen i det förväntade värdet för svarsvariabeln Y. Y förknippas med en ökning på en enhet i motsvarande prediktorvariabel, som håller alla andra prediktorer konstanta.

β0, intercept: Representerar det förväntade värdet för svarsvariabeln Y när alla prediktorvariabler ($x1, x2, ..., xp$) är noll.

B1, lutningen: Varje (var i=1,2.....,p) representerar förändringen i det förväntade värdet på Y för en ökning med en enhet i motsvarande prediktorvariabel (xi) som håller alla andra prediktorvariabler 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?

Svar: BIC hjälper till med modellval genom att balansera modellkomplexitet och god passform, hjälpa till att välja en modell som förklarar data väl utan att vara alltför komplex. Att enbart förlita sig på BIC garanterar dock inte en modells generaliserbarhet eller dess prestanda på nya data. Det är därför viktigt att dela upp data i trainings-, validerings- och testset.

Genom att dela upp data kan vi bedöma hur väl modellen generaliserar till nya data. Genom att utvärdera modeller på ett separat test set får vi en mer realistisk uppskattning av deras prestanda i verkliga scenarier och kan skydda oss mot overfitting.

6. Förklara algoritmen för ”Best subset selection”.

Svar: Den "Best Subset Selection” syftar till att identifiera den optimala kombinationen av prediktorer för en regressionsmodell med hjälp av följande steg:

- Det börjar med en nollmodell som inte innehåller några prediktorer.
- Då passar alla kombinationer av prediktorer (k) in i olika modeller (m1, m2...mp)
- Sedan väljs de bästa modellerna med lägst RSS (Residual Sum of Squares) eller största R-kvadrat.
- Välj sedan slutligen den enskilt bästa modellen från dessa modellers utvärdering på validerings set med hjälp av statistiska mått som prediktionsfel, AIC, BIC, Adjusted R squared eller CV.

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

Svar: Det betyder att ingen modell korrekt kan representera komplexiteten i det verkliga scenariot. Modeller är förenklingar av verkligheten och lyckas därför inte fånga alla dess aspekter. Trots deras brister kan modeller fortfarande ge värdefulla insikter, prediktioner eller förklaringar om den verkliga världen. Box försöker förmedla att istället för att sträva efter perfekta modeller, inser man att ofullkomliga modeller fortfarande kan vara värdefulla om de ger användbara insikter eller prediktioner.

7 Självutvärdering

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

Utmaningarna under det här projektet var att det blev väldigt förvirrande däremellan med så många modeller för jag ville experimentera med allt. Till slut började det bli vettigt när jag gjorde om allt några gånger för att riktigt förstå vad som händer.

2. **Vilket betyg du anser att du skall ha och varför.**

Jag anser att ha VG, eftersom jag har skapat ett robust arbetsflöde samtidigt som jag uppfyller alla villkor som krävs för provet.

References

- Taylor, S. (n.d.). Multiple Linear Regression. Corporate Finance Institute. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/
- IBM. (n.d.). Lasso Regression. Retrieved from https://www.ibm.com/topics/lasso-regression#:~:text=Lasso%20regression%20is%20a%20regularization,the%20accuracy%20of%20statistical%20models.
- Figure 2. Retrived from: https://medium.com/analytics-vidhya/new-aspects-to-consider-while-moving-from-simple-linear-regression-to-multiple-linear-regression-dad06b3449ff
- Figure 3: Retrieved from: https://www.statology.org/lasso-regression/
- Chat GPT