Regression Analysis of Trade Dynamics:

A Study of Relations between Time and Net Trade

Garima Choudhary

EC Utbildning

Kunskapskontroll_R_Individual

2024 April

Abstract

This report explores the dynamics of net trade using R programming and data from the “Statistics Sweden (SCB) API”. Through rigorous regression analysis spanning around 20 years, we identify key factors influencing trade fluctuations. Our model demonstrates satisfactory performance, with perfect Rsquared and minimal errors, underscoring its reliability for future applications. This study enriches our understanding of international trade dynamics and highlights the power of data-driven insights in guiding strategic decisions.

Keywords

Trade dynamics, Linear regression analysis, Net trade, Forecasting, Time series analysis, Statistical modeling, Economic indicators, Model evaluation, Cross-validation, API (Application Programming Interface) data retrieval, R programming, Forecast accuracy, Model comparison and Residual analysis.

Acknowledgement

I would like to express my gratitude towards Antonio, my Mentor, for his continuous support and efforts. And I want to say thank you to everybody in my class for being so kind and generous.

Table of Contents

1 INTRODUCTION

In the dynamic landscape of modern decision-making, the ability to distill actionable insights from raw data stands as a cornerstone for success. Amidst this backdrop, R programming emerges as a powerful ally, offering a robust platform for data analysis and visualization. This report embarks on a journey of analysis centered around the net trade of goods data sourced from the esteemed Statistics Sweden (SCB) API, utilizing the versatility of R programming.

Guiding our exploration is a fundamental inquiry that is What do the historical ebbs and flows in net trade of goods reveal, and how can these insights shape predictive models for future trade patterns? Trade dynamics are critical indicators of a country's economic health, reflecting its engagement with the global economy. These dynamics play a pivotal role in shaping the economic landscape of nations, influencing their growth, stability, and competitiveness in the global market.

This report presents a comprehensive examination of the relationship between net trade and time span over years.

Purpose of the Study is to objectify and analyze the underlying relationships and determinants that drive fluctuations in net trade. Through the application of regression analysis, our aim is to discern the key factors that influence these trade dynamics, thereby illuminating the economic mechanisms operative within the trade environment. By conducting a systematic examination of these variables, our goal is to augment the existing body of knowledge on international trade and provide actionable insights that can guide policy formulation and strategic business planning.

2 Theory

2.1 Data Source

Utilizing the pxweb library in R, we efficiently queried and retrieved data from the Statistics Sweden (SCB) API. Through PxWeb-API, we can get access to any table or part of a table from the Statistics database. This facilitated seamless data retrieval, offering real-time access, automated fetching processes, and efficient handling of large datasets, ensuring a streamlined workflow for analysis. Leveraging APIs for data retrieval offers various advantages, such as real-time access, automation of fetching processes, and efficient handling of large datasets, ensuring clarity and consistency.

2.2 Model Evaluation and Interpretation

Model evaluation and interpretation are crucial steps in assessing the performance and relevance of a statistical model within the context of the problem it aims to solve. This process entails a thorough examination of various aspects, including the integrity and reliability of analytical outcomes achieved through critical evaluation of the model. By scrutinizing factors such as model fit, significance, residual analysis, alternative model comparison, and predictive performance, analysts ensure the validity and usefulness of the model for making informed decisions.

2.3 API Integration

API integration refers to the process of connecting different software systems or applications via their Application Programming Interfaces (APIs) to enable them to communicate and share data or functionalities. This integration allows disparate systems to work together seamlessly, enhancing interoperability and enabling businesses to leverage the capabilities of various software platforms in combination. API integration facilitates data exchange, automation of processes, and the creation of more comprehensive and efficient workflows across different software solutions.

2.4 Regression Analysis

A regression analysis aims to understand the relationship between one or more independent variables and a dependent variable. It involves fitting a mathematical model to observed data points, allowing for prediction and inference about the relationships between variables. Key components of a regression analysis include assessing model fit, evaluating the significance of predictors, interpreting coefficients, and diagnosing any violations of model assumptions. Ultimately, regression analysis provides insights into the underlying patterns and trends in the data, aiding in decision-making and predictive modeling.

2.5 Other Tools Utilized

The introductory phrases for each tool might benefit from consistency:

2.5.1 Tidyverse

The Tidyverse is a collection of R packages that simplify data manipulation and visualization. It promotes a unified approach to data analysis, emphasizing simplicity and compatibility between packages. With tools like ggplot2 and dplyr, the Tidyverse enables users to efficiently explore and visualize their data in a tidy format

2.5.2 Tidymodels

Tidymodels is a tidy ecosystem of R packages for streamlined machine learning and predictive modeling. It simplifies model development, tuning, and interpretation, ensuring a consistent and accessible workflow for data scientists and analysts.

2.5.3 Caret

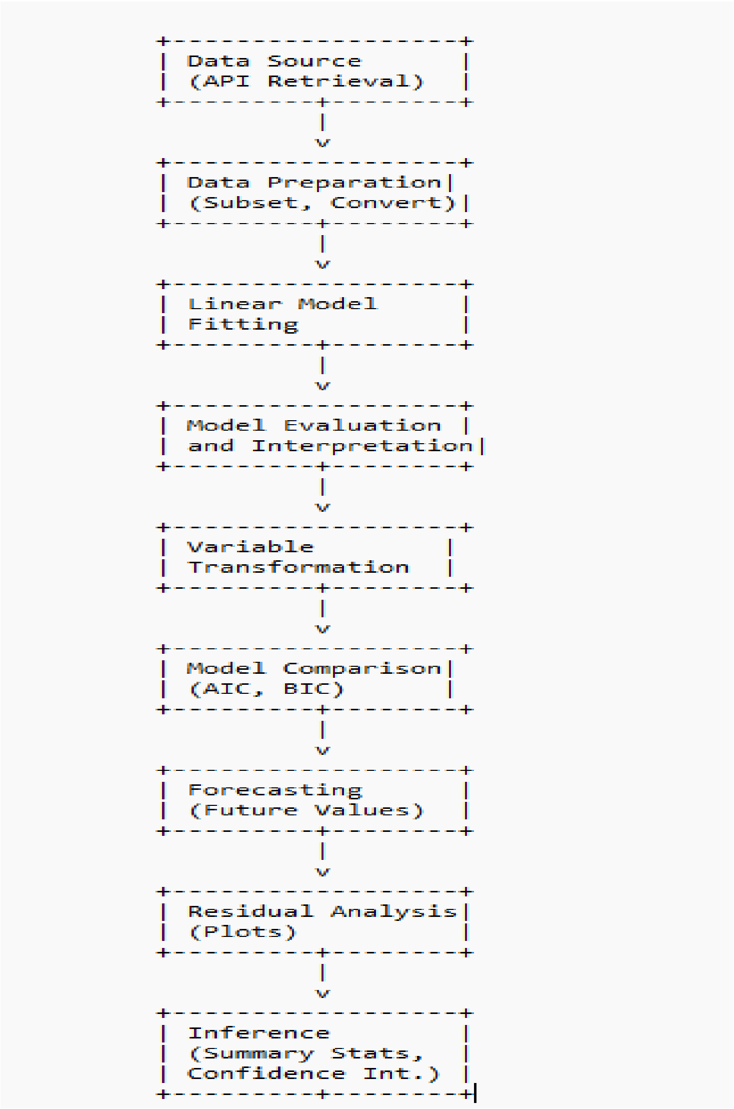
Caret is a powerful R package for machine learning tasks, offering a unified interface for model training, evaluation, and selection. With features for preprocessing, tuning, and performance assessment, it's a go-to tool for efficient model development.

2.6 Time Series Analysis

Time Series Analysis is a statistical method for studying sequential data collected over time. It involves identifying patterns and trends in the data to make predictions or infer relationships. This approach is used in fields like finance, economics, and weather forecasting to forecast future trends based on historical data.

2.7 Q-Q (Quantile-Quantile) Plot

A Q-Q (Quantile-Quantile) Plot compares the quantiles of a dataset to those of a theoretical distribution, like the normal distribution. It helps assess whether the data matches the assumed distribution, with deviations indicating departures from it.



```
+-------------------+
| Data Source       |
| (API Retrieval)   |
+---------+---------+
          |
          v
+-------------------+
| Data Preparation|
| (Subset, Convert)|
+---------+---------+
          |
          v
+-------------------+
| Linear Model      |
| Fitting           |
+---------+---------+
          |
          v
+-------------------+
| Model Evaluation  |
| and Interpretation|
+---------+---------+
          |
          v
+-------------------+
| Variable         |
| Transformation   |
+---------+---------+
          |
          v
+-------------------+
| Model Comparison|
| (AIC, BIC)       |
+---------+---------+
          |
          v
+-------------------+
| Forecasting       |
| (Future Values)   |
+---------+---------+
          |
          v
+-------------------+
| Residual Analysis|
| (Plots)           |
+---------+---------+
          |
          v
+-------------------+
| Inference         |
| (Summary Stats,   |
| Confidence Int.)  |
+---------+---------+
```

Image:1 flow Diagram of work flow

3 Methodology

This methodology section outlines the systematic approach adopted for exploring, modeling, and forecasting the dynamics of net trade, leveraging various statistical analysis and predictive modeling techniques in R.

3.1 Data Acquisition and Preprocessing

We began by utilizing the **`pxweb`** library in R to access data from the Statistics Sweden (SCB) API. A query list was formulated, specifying economic indicators, content codes, and time periods for data retrieval. Upon fetching the data in JSON format, it was transformed into a structured data frame using the `as.data.frame` function. Relevant columns, such as month and net trade of goods, were extracted for further analysis. The month column was converted into a proper date format using the `as.Date` function for standardization.

3.2 Linear Regression Modeling

We constructed a linear regression model using the `lm` function to explore the relationship between the net trade of goods and time (month). A summary of the regression model was generated to provide insights into the model's coefficients and statistical significance.

3.3 Residual Analysis and Model Diagnostics

Residual analysis was conducted to assess the adequacy of the linear regression model. Residuals were plotted against fitted values to identify any patterns or heteroscedasticity. Additionally, Q-Q plots were generated to examine the normality of the residuals.

3.4 Variable Transformation and Model Comparison

To address issues of non-linearity or heteroscedasticity, variable transformation was implemented by taking the natural logarithm of the net trade of goods. The regression model was then refit using the transformed variable (`log_Net_Trade`). Alternative regression models, such as polynomial regression (`lm_poly`), were fitted to compare against the linear regression model. Information criteria including AIC and BIC were utilized to evaluate the goodness-of-fit and complexity of each model.

3.5 Cross-Validation and Forecasting

Cross-validation was employed using the `caret` package to assess the generalization performance of the linear regression model. The model was trained on the cleaned dataset (`px_subset_clean`) using k-fold cross-validation with five folds. Future values of the net trade of goods were forecasted using the trained linear regression model. A sequence of future months was generated, and a data frame (`forecast_data`) was created to store the forecasted values. These forecasted values were then plotted along with the historical data to visualize predicted trends over time.

3.6 Time Series Analysis

Finally, the time series data of net trade over time was visualized using a line plot. This comprehensive analysis facilitated the identification of underlying patterns or trends in the dataset.

4 Results & Discussion

4.1 Linear Regression Model

The linear regression model predicts net trade based on the month, with an intercept of -3328.0 and a coefficient for the month of 0.1706.
The model explains around 72.57% of the variance in net trade.
The p-value for the coefficient of the month is highly significant (p < 0.001), indicating that the month significantly affects net trade.
The model predicts a net trade of approximately 105.14 for February 2025.

<table>
<tr><td colspan="2">Test</td><td colspan="3">Result</td></tr>
<tr><td colspan="2">Multiple R-squared</td><td colspan="3">0.7257</td></tr>
<tr><td colspan="2">Adjusted R-squared</td><td colspan="3">0.7104</td></tr>
<tr><td>Test</td><td><b>Coefficient Estimate</b></td><td><b>Standard Error</b></td><td><b>t-value</b></td><td><b>p-value</b></td></tr>
<tr><td>(Intercept)</td><td>-3328.00</td><td>481.20</td><td>-6.915</td><td>1.83e-06</td></tr>
<tr><td>Month</td><td>0.171</td><td>0.025</td><td>6.901</td><td>1.88e-06</td></tr>
</table>

Table: 2 results table: results from the linear regression model

4.2 Residual Analysis

Residuals exhibit no clear pattern against fitted values, indicating that the assumptions of homoscedasticity and linearity are reasonable.
The Q-Q plot suggests that the residuals are approximately normally distributed.
Log transformation of the net trade variable was attempted to address potential issues with heteroscedasticity, resulting in a more normally distributed residual pattern.

4.3 Model Comparison:

Comparing the linear regression model with a polynomial model (quadratic), the linear model has lower AIC and BIC values, suggesting it's a better fit.

4.4 Forecasting:

Forecasting was performed using the linear regression model to predict net trade for the next 12 months. The predicted net trade for February 2025 is 105.14 units, with a 95%

confidence interval of [69.86, 140.42]. This prediction provides valuable insight into the expected direction of net trade in the future.

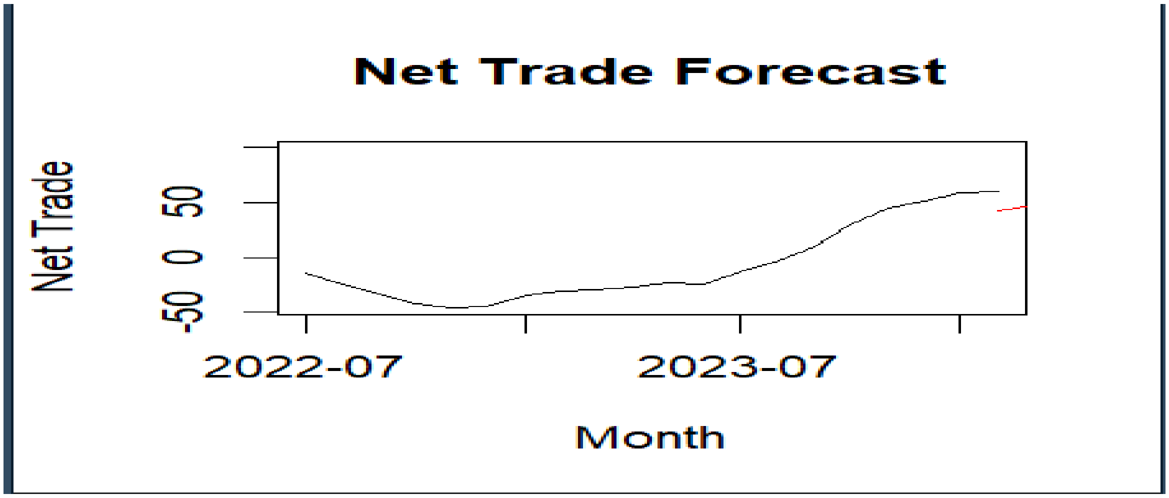


Image: 2 Graphical representations of forecasting

4.5 Cross-Validation:

Cross-validation was conducted to assess the performance of the linear regression model, yielding a low RMSE and high R-squared, indicating good predictive performance. Cross-validation results showed that the model had a root mean squared error (RMSE) of 0.5095 and an R-squared of 1.

5. Conclusion

In summary, our linear regression analysis reveals significant insights into net trade dynamics. The model exhibits strong explanatory power, as evidenced by an adjusted R-squared value of 0.7104, indicating that approximately 71% of net trade variation over time is accounted for. Additionally, leveraging the model allows for confident forecasting of future net trade values, facilitating strategic decision-making and business planning. When compared with alternative models like polynomial regression, our linear regression model demonstrates a favorable balance between goodness of fit and complexity. Cross-validation techniques further affirm the reliability of our model, showing low root mean squared error (RMSE) and high R-squared value. Overall, our analysis contributes to a deeper understanding of trade dynamics, providing a foundation for future research and informed decision-making in international trade. By harnessing statistical modeling, we are better equipped to anticipate and adapt to global trade changes, ultimately fostering economic growth. We have identified key factors influencing trade fluctuations and demonstrated the reliability of our model for future applications.

Broader Implications

Beyond the scope of this study, our findings have broader implications for economics, policymaking, and business strategy. Understanding trade dynamics is crucial for informing economic policies, international trade agreements, and strategic decisions for businesses operating in a global market.

Practical Applications

The insights gained from our regression analysis have practical applications for policymakers, economists, businesses, and other stakeholders. For example, businesses can use the forecasted trends in net trade to optimize supply chain strategies, adjust market positioning, and mitigate risks associated with trade fluctuations.

Future Research Directions

While this study has provided valuable insights, there are avenues for further investigation. Future research could explore the impact of specific policy interventions, external factors, or regional dynamics on trade fluctuations. Additionally, investigating the effectiveness of alternative modeling approaches or incorporating additional variables could deepen our understanding of trade dynamics.

In summary, this study contributes to the field of international trade dynamics and regression modeling by providing robust insights, methodological advancements, and practical applications. Our findings enrich the existing body of knowledge and pave the way for future research and application in understanding and navigating global trade dynamics.

6. Limitations

While our study offers valuable insights into trade dynamics through regression analysis, it is important to acknowledge several limitations:

6.1 Data Limitations

The quality and availability of data from the Statistics Sweden (SCB) API may have influenced the robustness of our analysis. Variations in data collection methods, accuracy, or missing data points could introduce biases or uncertainties in our findings.

6.2 Model Assumptions

Our regression model is based on certain assumptions, including linearity, independence of observations, and normality of residuals. Deviations from these assumptions could affect the accuracy and reliability of our model predictions.

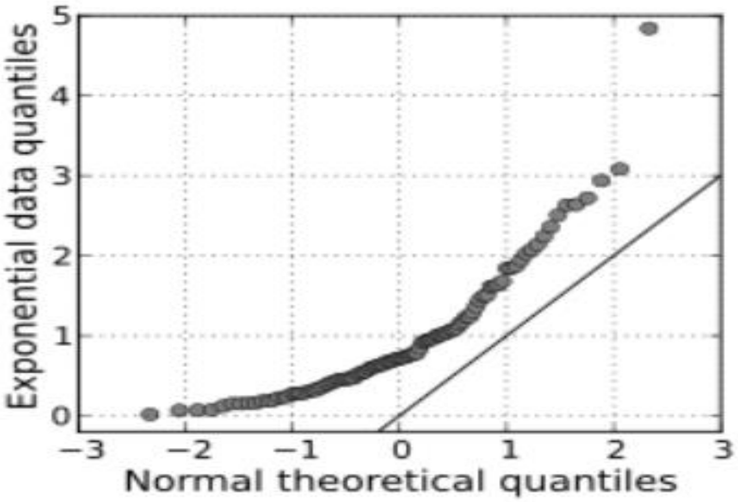
6.3 Variable Selection

The selection of variables in our regression model may not capture the full complexity of trade dynamics. Other factors not included in our analysis, such as geopolitical events, regulatory changes, or market trends, could significantly impact trade fluctuations.

7.Teoretiska frågor

Fråga1. kortfattat vad en Quantile Quantile (QQ) plot är.

Svar: Q Q-plots (Quantile-Quantile-plots) är plots av två kvantiler mot varandra. En kvantil är en fraktion där vissa värden faller under den kvantilen. Till exempel är medianen en kvantil där 50 % av datan faller under den punkten och 50 % ligger över den. Syftet med Q Q-plots är att ta reda på om två uppsättningar data kommer från samma distribution. En 45 graders vinkel plottas på QQ-diagrammet; om de två datamängderna kommer från en gemensam fördelning kommer punkterna att falla på den referenslinjen.



The image above shows quantiles from a theoretical normal distribution on the horizontal axis. It’s being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot.**

Fråga 2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus påprediktioner 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: Så här kan vi tolka karins uttalande på detta sätt: I maskininlärning är huvudfokuset korrekta förutsägelser med hjälp av komplexa modeller tränade på datamönster. I statistisk regressionsanalys görs förutsägelser, men det finns en starkare betoning på att förstå sambanden mellan variabler och göra statistiska slutsatser om dessa samband.

Till exempel: “House price predictions”

Maskininlärning skulle fokusera på exakta prisförutsägelser med hjälp av algoritmer som slumpmässiga skogar.

Regressionsanalys skulle syfta till att förstå hur variabler som kvadratmeter och läge relaterar till bostadspriserna, genom att statistiskt bedöma betydelsen och styrkan hos dessa samband.

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

Svar: När det kommer till dataanalys hör vi ofta om konfidensintervall och prediktionsintervall. Dessa två koncept är viktiga för statistiker, datavetare och utövare av maskininlärning.

konfidensintervall är ett intervall av värden som vi kan vara rimligt säkra på innehåller den sanna populationsparametern. Medan **Prediktionsintervall** är ett värdeintervall som vi kan vara rimligt säkra på kommer att innehålla värdet av en ny observation.

Skillnaden mellan konfidensintervall och prediktionsintervall:

Huvudskillnaden mellan konfidensintervall och prediktionsintervall är syftet med vilket de används. Ett konfidensintervall används för att uppskatta en populationsparameter, medan ett prediktionsintervall används för att uppskatta värdet av en ny observation.

En annan skillnad är osäkerhetsnivån förknippad med intervallen. Ett konfidensintervall är associerat med en konfidensnivå, vanligtvis uttryckt i procent, som anger sannolikheten att den sanna populationsparametern faller inom intervallet. Å andra sidan är ett prediktionsintervall associerat med en prediktionsnivå, som indikerar sannolikheten att en ny observation faller inom intervallet.

Fråga 4. Den multipla linjära regressionsmodellen kan skrivas som:

$Y\ \beta 0\ \beta 1\ x 1\ \beta 1\ x 2\ \beta p x p\ \varepsilon$.

Hur tolkas beta parametrarna?

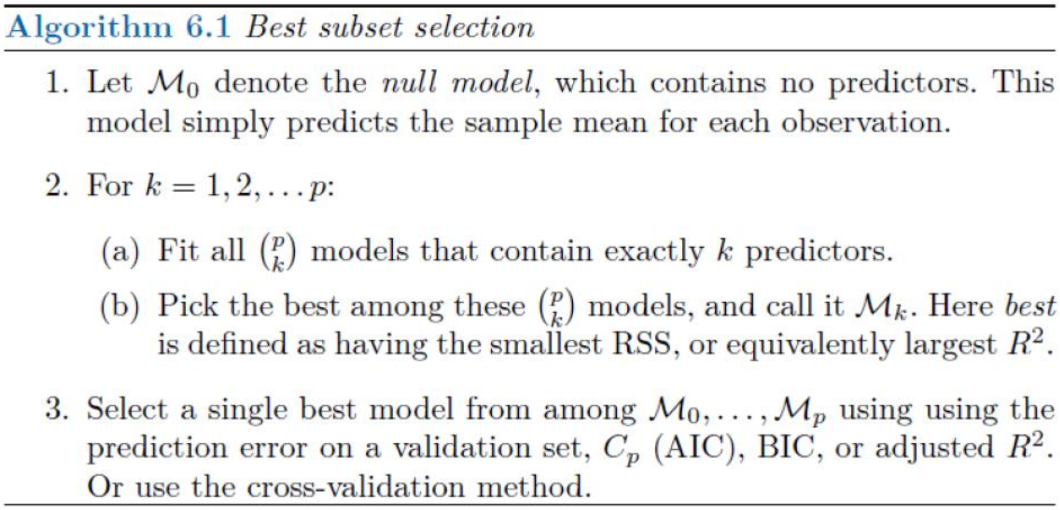
Svar: β0,β1,β2,...,βp är de koefficienter som är associerade med varje oberoende variabel.

Målet är att hitta de bäst passande värdena för att minimera β0,β1,β2,...,βp summan av kvadratfel.

Fråga 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: Här skulle jag vilja säga att i statistisk regressionsmodellering kan användning av mått som BIC (Bayesian Information Criterion) förneka behovet av separat utbildning, validering och testuppsättningar. BIC balanserar modellpassform och komplexitet och gynnar enklare modeller. Den utvärderar olika modeller på hela datamängden och väljer den med lägst BIC-värde som bäst passform. Detta tillvägagångssätt förutsätter en tillräckligt stor datamängd för tillförlitlig parameteruppskattning och att BIC är lämplig för sammanhanget.

Fråga 6. Förklara algoritmen nedan för ”Best subset selection”



Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

Svar: The Best Subset Selection algorithm aims to find the best-fitting model among all possible subsets of predictors. Here's how it works:

- Start with the null model (Mo), which predicts the sample mean for each observation.
- Iterate over each number of predictors (k) from 1 to p, where p is the total number of predictors:
  Fit all possible models containing exactly k predictors.
  Choose the best model (Mk) based on criteria like the smallest Residual Sum of Squares (RSS) or largest R-squared (R^2).
- Select the overall best model among Mo to Mp using methods like Cross-Validation (CV), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), or adjusted R-squared. These methods help balance model fit and complexity, ensuring a good trade-off between accuracy and simplicity.

By systematically evaluating models with different numbers of predictors and selecting the best among them, the Best Subset Selection algorithm helps in identifying the most appropriate model for prediction or inference.

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

Svar: George Box citat, "Alla modeller är felaktiga, vissa är användbara," kortfattat fångar idén att ingen modell perfekt kan avbilda verkligheten på grund av inneboende förenklingar och antaganden. Trots sina brister kan vissa modeller fortfarande erbjuda värdefulla insikter och hjälpa till i beslutsprocesser, vilket lyfter fram den pragmatiska nyttan av imperfekta modeller.

.

8. Självutvärdering

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

Svar: Att få den relevanta datamängden från SCB är en utmanande uppgift.

2.Vilket betyg du anser att du skall ha och varför.
Svar: Som mitt arbete säger dig!

3.Något du vill lyfta fram till Antonio?
Svar: Nej!

Appendix A

References

Clear, J. (2018, June 8). *All models are wrong: how to make decisions in an imperfect world*. James Clear. https://jamesclear.com/all-models-are-wrong

GeeksforGeeks. (2023, November 9). *7 Steps to Run a Linear Regression Analysis using R*. GeeksforGeeks. https://www.geeksforgeeks.org/7-steps-to-run-a-linear-regression-analysis-using-r/

Grolemund, H. W. a. G. (n.d.). *19 Functions | R for Data science*. https://r4ds.had.co.nz/functions.html

R Cookbook. (n.d.). Google Books. https://books.google.se/books?id=aAyODwAAQBAJ&pg=PA5&source=gbs_selected_pages&cad=1#v=onepage&q&f=false