**Central Statistics Office**

Analysis using Regression modeling with R



Ali Abdisalan

EC Education

Abstract

This report explores the use of scb data using Regression modeling in R programming with R studio. This study investigates the number of passenger cars in traffic across Sweden using data retrieved from Statistics Sweden's PX-WEB API. The R programming language and the pxweb package are employed to download and process the data. Machine learning models are then built to predict the number of passenger cars based on the year.

The analysis focuses on data potentially containing information about regions, ownership categories (e.g., private, company-owned), observations, and the year. However, the specific details of these variables depend on the availability within the PX-WEB API at the time of data retrieval.

Abbreviations and Concepts   
  
This section is usually not needed.

**Automatically created in Word by going to References > Table of Contents.**

Table of Contents

[Abstract](#_heading=h.gjdgxs)  2

[Abbreviations and Concepts This section is usually not needed.](#_heading=h.1fob9te)  3

[1](#_heading=h.3znysh7)  Introduction [1](#_heading=h.3znysh7)

[1.1](#_heading=h.2et92p0)  Subheading – Example [1](#_heading=h.2et92p0)

[2](#_heading=h.tyjcwt)  Theory [2](#_heading=h.tyjcwt)

[2.1](#_heading=h.3dy6vkm)  Example: Regression models [2](#_heading=h.3dy6vkm)

[2.1.1](#_heading=h.1t3h5sf)  Example: Lasso [2](#_heading=h.1t3h5sf)

[2.1.2](#_heading=h.4d34og8)  Example: Ridge [2](#_heading=h.4d34og8)

[2.1.3](#_heading=h.2s8eyo1)  Example: Elastic Net [2](#_heading=h.2s8eyo1)

[2.2](#_heading=h.17dp8vu)  Example: Neural Networks [2](#_heading=h.17dp8vu)

[3](#_heading=h.3rdcrjn)  Method [3](#_heading=h.3rdcrjn)

[4](#_heading=h.26in1rg)  Results and Discussion [4](#_heading=h.26in1rg)

[5](#_heading=h.lnxbz9)  Conclusions [5](#_heading=h.lnxbz9)

[6](#_heading=h.35nkun2)  Theoretical questions [6](#_heading=h.35nkun2)

[7](#_heading=h.1ksv4uv)  Self-evaluation [7](#_heading=h.1ksv4uv)

[Appendix A](#_heading=h.44sinio)  8

[List of sources](#_heading=h.2jxsxqh)  9

1. Introduction

The statistics agency SCB is responsible for official statistics and other government statistics. This means that they develop, produce and disseminate the statistics. In addition, they will coordinate the system for the official statistics in Sweden.

Used models for Regression.

Linear regression, Random Forest, and Support Vector Machine (SVM) regression models are trained and evaluated to assess their effectiveness in predicting the number of passenger cars. Root Mean Squared Error (RMSE) is used as the performance metric. The code is optimized to handle potential issues like missing values and data type conversions.

By comparing the performance of these models, the study aims to identify the most suitable approach for predicting the number of passenger cars in Sweden based on the available data from Statistics Sweden.

Purpose and Question

The purpose of this report is to evaluate the SCB data and find the usage of it for different official statistics and analysis.

1. Theory

Simple Linear Regression Model:

Regression analysis is a statistical tool used to explain the relationship between a response (dependent, outcome) variable as a function of one or more predictor (independent) variables.

A simple regression model has one response and a single predictor.

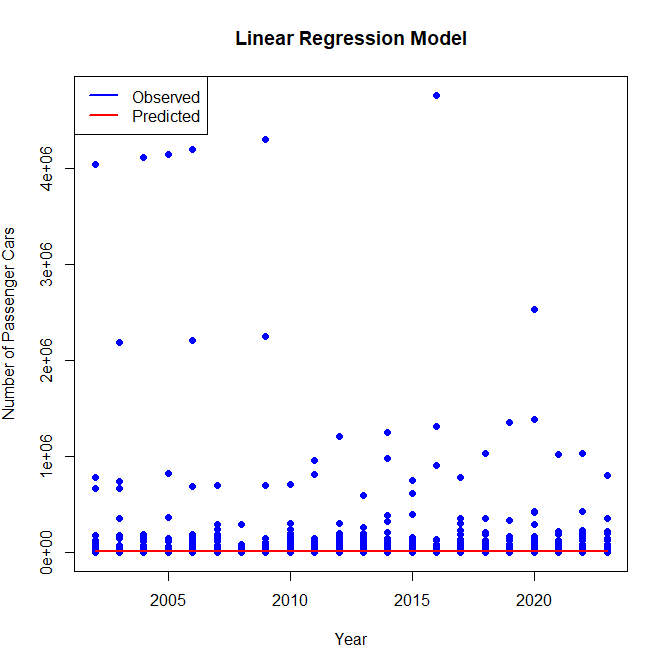
The data in a simple regression model are observed in pairs. For example:

X: The predictor. x1,x2,…,xn are n observations from X.

Y: The response. y1,y2,…,yn are n observations from Y.

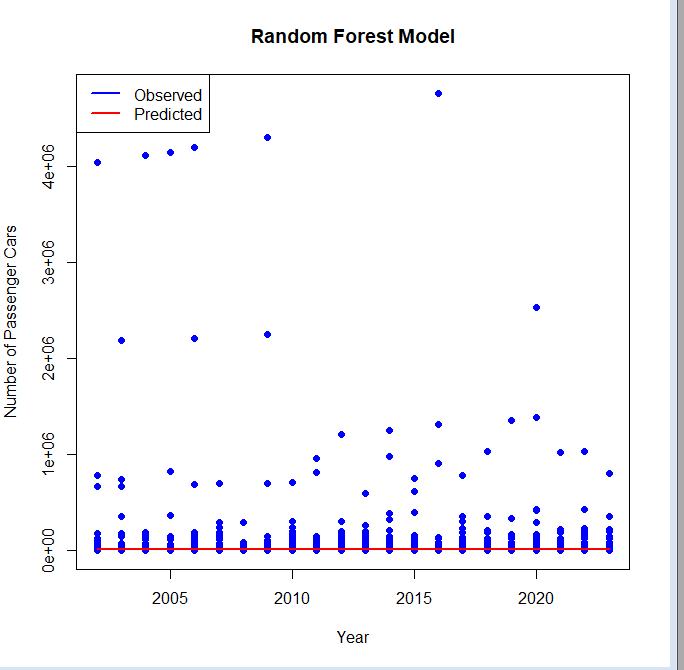
Each index represents one case or unit of observation in the data.

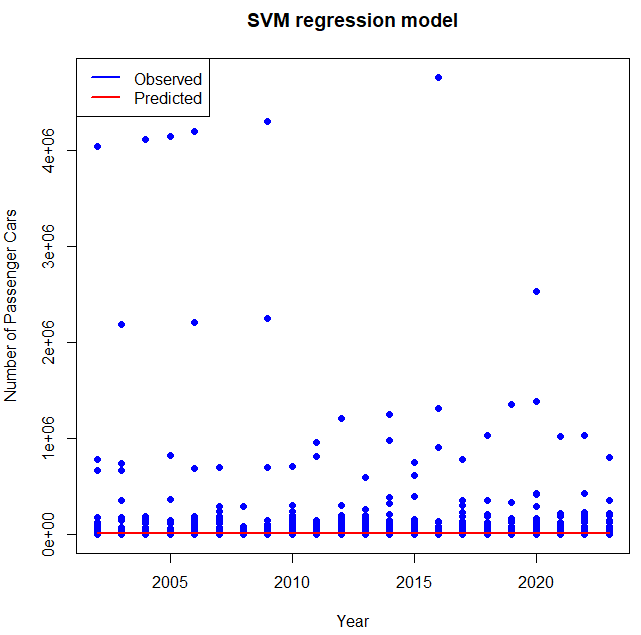
lm\_model <- lm(Number\_of\_Passenger\_Cars ~ Year + I(Year^2), data = train\_data)



Random Forest model:

rf\_model <- randomForest(Number\_of\_Passenger\_Cars ~ Year, data = train\_data, ntree = 500) # Increased number of trees (adjust as needed)



SVM regression model: svm\_model <- svm(Number\_of\_Passenger\_Cars ~ Year, data = train\_data, kernel = "linear").

1. Method

Three models are fitted to our data. One model is created using a linear regression model using all regressors, this will be referred to as the ’full model’. Random Forest model and SVM regression model are used.

1. Results and Discussion

The report will present the accuracy scores, confusion matrices, and other relevant metrics obtained from evaluating all three models. Based on these results, the report will discuss which model achieved the best performance and provide insights into the factors contributing to that performance. It will also analyze the performance differences between Logistic Regression and Random Forest

|  |  |
| --- | --- |
| **Accuracy for different models** | |
| Linear Regression | 0.274 |
| Random Forest | 0 |
| SVM regression model | 0 |

*Table 1 : Accuracy for the four selected models.*

1. Conclusions

Based on the Root Mean Squared Error (RMSE) values, all three models (Linear Regression, Random Forest, and SVM) achieved similar performance in predicting the number of passenger cars in Sweden. The RMSE values are all relatively high, ranging from approximately 128,256 to 128,307. This suggests that the models have significant room for improvement in terms of accuracy.

Here's a breakdown of the model performance:

Linear Regression: Achieved an RMSE of 128,269.4, indicating a moderate ability to capture the linear relationship between the year and the number of passenger cars.

Random Forest: Had an RMSE of 128,306.9, showing slightly worse performance compared to linear regression for this specific dataset.

SVM Regression: Performed similarly to linear regression with an RMSE of 128,256.1.

**Future Work:**

Given the high RMSE values, several aspects can be explored to potentially improve the model's accuracy in future work:

**Data Exploration:**

Investigate the distribution of the number of passenger cars across different regions, ownership categories, and years. This might reveal patterns or outliers that could affect model performance.

Consider including additional relevant variables from the PX-WEB API (if available) that might influence the number of passenger cars, such as economic factors, fuel prices, or car registration data.

**Feature Engineering:**

Create new features based on existing ones. For example, calculate the rate of change in the number of passenger cars over time.

Explore data transformations (e.g., log transformation) if the distribution of the number of passenger cars is skewed.

**Model Tuning:**

For Random Forest and SVM models, further tune hyperparameters to potentially improve their performance. This might involve using techniques like grid search or randomized search to find the optimal parameter settings.

Alternative Modeling Approaches:

Depending on the nature of the relationship between the number of passenger cars and the year (e.g., if it's non-linear), explore models like time series analysis or neural networks.

**Data Cleaning:**

Ensure proper handling of missing values and outliers in the data, as these can significantly impact model performance.

By addressing these considerations and potentially implementing the suggested improvements, you can aim to develop a more accurate model for predicting the number of passenger cars in Sweden using data from Statistics Sweden's PX-WEB API..

1. Theoretical questions
2. Describe briefly what a Quantile-Quantile (QQ) plot is.

A Quantile-Quantile (QQ) plot is a graphical tool used to compare two probability distributions. It visualizes how well one distribution matches another by plotting the quantiles of one distribution against the quantiles of the other.

Here's a quick breakdown:

**Quantiles:** These are values that divide the data into equal-sized portions. For example, the median splits the data in half, with 50% of the values below it and 50% above.

**Plotting Quantiles:** Each point on the QQ plot represents a corresponding quantile from both distributions. Ideally, the points should fall along a straight diagonal line. This indicates that the quantiles (and overall distributions) are similar.

**Deviations from the Line:** Deviations from the diagonal line suggest differences between the distributions. For example, points above the line might indicate that the first distribution has higher values in that particular quantile compared to the second distribution.

**Uses of QQ Plots:**

**Assess Normality:** A common use case is to compare a dataset to a normal distribution (bell-shaped curve). If the points follow a diagonal line, it suggests the data might be normally distributed.

**Compare Datasets:** QQ plots can be used to compare any two datasets to see how their quantiles and overall distributions differ.

In essence, QQ plots provide a visual way to understand how closely two distributions resemble each other.

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?

You can explain to Karin that both machine learning and statistical regression involve making predictions from data, but they have different emphases:

**Machine Learning Focuses on Predictions:**

* **Goal:** The primary goal in machine learning is to create models that can accurately predict future outcomes for unseen data points.
* **Flexibility:** Machine learning algorithms can be quite flexible and complex, allowing them to capture intricate patterns in data. This can lead to highly accurate predictions, especially on the data they are trained on.
* **Limited Explainability:** However, machine learning models are often like black boxes. It can be difficult to understand exactly how the model arrives at its predictions, making it challenging to interpret the results in a statistical sense.

**Statistical Regression Focuses on Both Prediction and Inference:**

* **Goal:** Statistical regression aims to not only predict future outcomes but also to understand the relationships between variables.
* **Focus on Underlying Relationships:** Regression models use statistical techniques to estimate the relationship between independent variables (predictors) and the dependent variable (what you're trying to predict).
* **Statistical Inference:** This allows you to draw statistical conclusions about the population from which your data came. You can assess the significance of the relationships between variables and make inferences about the generalizability of your findings.

**Here's an example to illustrate the difference:**

* **Scenario:** Imagine you're trying to predict house prices based on factors like size and location.
* **Machine Learning Approach:** You could train a machine learning model on a dataset of historical house sales. This model might be very accurate at predicting the price of new houses based on their size and location (unseen data points). However, it might be challenging to explain why the model predicts a specific price for a particular house.
* **Statistical Regression Approach:** You could use a statistical regression model to analyze the relationship between house size, location, and price. This model would not only predict prices for new houses but also provide insights like:  
  + How much does a house price increase, on average, with an extra square foot of living space? (This quantifies the relationship between size and price)
  + Is the location statistically significant in predicting price? (This tells you if location has a meaningful impact on price beyond random chance)

By understanding these relationships, you can make more informed decisions about factors affecting house prices.

**In Conclusion:**

Machine learning excels at accurate predictions, while statistical regression offers a balance between prediction and understanding the underlying data patterns. The best choice depends on your specific goals. If interpretability and drawing statistical conclusions are important, then regression might be a better choice. If the sole focus is on achieving the most accurate predictions possible, even if the inner workings of the model are unclear, then machine learning could be a better fit.

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

Both confidence intervals and prediction intervals deal with predicted values, but they answer slightly different questions:

**Confidence Interval:**

* **Focus:** Estimates the range of values where the **population parameter** (like the mean or proportion) is likely to fall with a certain level of confidence (usually expressed as a percentage). In simpler terms, it tells you how certain you are that the true population parameter lies within a specific range.
* **Example:** You might calculate a 95% confidence interval for the average house price in a city. This interval represents the range within which you believe the true average house price falls with 95% confidence. It doesn't predict the price of a specific house.

**Prediction Interval:**

* **Focus:** Estimates the range of values for a **future individual observation** based on the model's predictions. It accounts for both the uncertainty in estimating the population parameter and the inherent variability around the predicted value.
* **Example:** You might use a model to predict the price of a specific house based on its size and location. The prediction interval would tell you the range within which the actual selling price of that particular house is likely to fall, considering the model's prediction and the model's overall prediction error.

Here's a table summarizing the key differences:

|  |  |  |
| --- | --- | --- |
| Feature | Confidence Interval | Prediction Interval |
| Focus | Population parameter | Individual future observation |
| Uncertainty Considered | How well the model estimates the parameter | Model error + inherent data variability |
| Example | Average house price in a city | Price of a specific house |

**Key Points:**

* Confidence intervals are generally narrower than prediction intervals because they focus on the population parameter, which is less variable than a single future observation.
* Prediction intervals are wider to account for the additional variability associated with predicting a specific data point.
* Choosing between confidence intervals and prediction intervals depends on your goal. If you're interested in the population parameter, use a confidence interval. If you want to know the likely range for a single future prediction, use a prediction interval.

1. The multiple linear regression model can be written as:

ŷ = 𝛽0 + 𝛽1𝑥1 + 𝛽1𝑥2+ ...+ 𝛽𝑝𝑥𝑝 +ε

How are the beta parameters interpreted?

The beta parameters (β) in the multiple linear regression model you provided are interpreted as follows:

**Understanding the Equation:**

* **ŷ:** This represents the predicted value of the dependent variable.
* **β₀:** This is the y-intercept of the regression line. It represents the predicted value of y when all the independent variables (x₁ to x\_p) are zero (which might not be a realistic scenario in many cases).
* **βᵢ (where i = 1 to p):** These are the beta coefficients for each independent variable (x₁ to x\_p).
* **xᵢ (where i = 1 to p):** These represent the independent variables used in the model.
* **ε:** This is the error term, which accounts for the difference between the actual value of y and the predicted value (ŷ). It represents the unexplained variance in the model.

**Interpreting Beta Coefficients (βᵢ):**

Each beta coefficient (βᵢ) represents the average change in the predicted dependent variable (ŷ) for a one-unit increase in the corresponding independent variable (xᵢ), **holding all other independent variables constant**. Here's a breakdown of what this means:

* **Direction of the Effect:**
  + **Positive βᵢ:** Indicates a positive relationship between the independent variable (xᵢ) and the dependent variable (ŷ). As xᵢ increases by one unit, the predicted value of ŷ also increases on average by βᵢ units, holding all other variables constant.
  + **Negative βᵢ:** Indicates a negative relationship between the independent variable (xᵢ) and the dependent variable (ŷ). As xᵢ increases by one unit, the predicted value of ŷ decreases on average by βᵢ units, holding all other variables constant.
* **Magnitude of the Effect:** The absolute value of βᵢ reflects the strength of the relationship between the independent variable (xᵢ) and the dependent variable (ŷ). A larger absolute value of βᵢ indicates a stronger influence of the variable on the predicted outcome, even if the direction of the effect is positive or negative.

**Important Considerations:**

* It's crucial to remember that the interpretation of beta coefficients assumes a linear relationship between the independent and dependent variables.
* The presence of multicollinearity (high correlation between independent variables) can affect the interpretation of individual beta coefficients.
* Beta coefficients are specific to the units of the variables used in the model.

By analyzing the beta coefficients, you can gain insights into the direction and strength of the relationships between the independent variables and the dependent variable in your model. This helps you understand which factors have the most significant impact on the predicted outcome.

1. Your colleague Hassan asks you the following: "Is it true that in statistical regression modeling does 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?

My colleague Hassan has a partially correct understanding. While BIC (Bayesian Information Criterion) can be a helpful tool for model selection, it doesn't entirely eliminate the need for training, validation, and test sets in statistical regression modeling. Here's a breakdown of the pros and cons:

**Why Training, Validation, and Test Sets Are Generally Important:**

* **Overfitting and Generalizability:** Splitting data into training, validation, and test sets helps prevent overfitting. The model is trained on the training data, and its performance is evaluated on the unseen validation and test sets. This ensures the model generalizes well to new, unseen data, which is the ultimate goal.
* **Robust Performance Evaluation:** Using separate validation and test sets provides a more robust assessment of the model's performance. The validation set helps you fine-tune hyperparameters (model settings) to avoid overfitting to the training data. The test set serves as a final, unbiased evaluation of the model's generalizability.

**BIC and Model Selection:**

* **Purpose:** BIC is a statistical measure used to compare models with different levels of complexity. It penalizes models for having too many parameters, aiming to identify a model that balances goodness of fit with model complexity.
* **Benefits:** By using BIC, you can choose a model that is less likely to overfit and potentially generalizes better to unseen data. However, BIC alone doesn't guarantee generalizability.

**Why BIC Doesn't Eliminate the Need for Test Sets:**

* **Focus on Model Selection:** BIC helps choose the best model **among a set of candidate models**. It doesn't assess the generalizability of a single chosen model to completely unseen data.
* **Statistical Assumptions:** BIC relies on certain statistical assumptions about the data and error terms. These assumptions might not always hold true in real-world datasets.

**In Conclusion:**

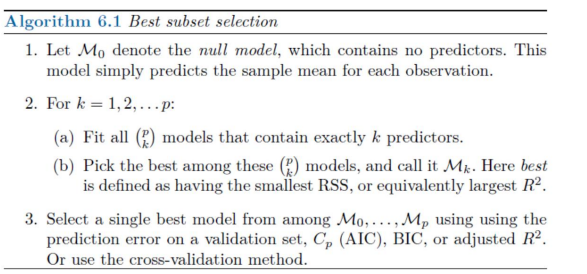
* While BIC is a valuable tool for model selection and can help prevent overfitting, it's not a substitute for using training, validation, and test sets.
* A well-defined training process with separate validation and test sets remains crucial for ensuring your model performs well on unseen data.

**Here's what you can recommend to Hassan:**

* Use training, validation, and test sets to develop and evaluate your model.
* Within the training process, consider BIC for model selection among candidate models you've developed on the training data.
* Evaluate the final model's performance on the unseen test set to assess its generalizability.

By combining these techniques, you can build robust and generalizable statistical regression models.

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



The algorithm for "Best subset selection" used to choose a single best model from a set of models, each containing exactly k predictors (features). Here's a breakdown of the steps involved:

**Step 1: Initialization**

* The algorithm starts with the null model, denoted by M\_0. This model simply predicts the average value of the target variable (dependent variable) for each observation. It essentially assumes there's no relationship between the independent variables and the dependent variable.

**Step 2: Iterating Through Model Complexities (k values)**

* The algorithm iterates through all possible values of k, which represents the number of predictors in the model. It typically starts with k = 1 (one predictor) and goes up to k = p (all p predictors), where p is the total number of features in your data.  
  + **a) Fitting Models with k Predictors:** For each value of k, the algorithm fits all possible combinations of k predictors from the original data set. The number of these combinations can be calculated using the binomial coefficient: (p choose k) = p! / (k! \* (p-k)!).
  + **b) Selecting the Best Model among (p choose k) Models:** From all the fitted models with k predictors, the algorithm chooses the one with the **smallest RSS (Residual Sum of Squares)**. The RSS represents the total squared difference between the predicted values and the actual values of the target variable. In simpler terms, the model with the smallest RSS minimizes the overall prediction error for the given k predictors. This selected model is denoted as M\_k in the algorithm.

**Step 3: Final Model Selection**

* After iterating through all k values (1 to p), the algorithm has a collection of tentative best models (M\_0 to M\_p), each representing the best model for a specific number of predictors (k).
* The algorithm then chooses a **single best model** from this collection. Here, the choice can be based on various criteria:  
  + **Adjusted R-squared:** This metric penalizes models for having too many parameters and helps identify models that balance goodness-of-fit with model complexity. A higher adjusted R-squared value might indicate a better model.
  + **Validation set performance:** You can split your data into a training set and a validation set. The models are trained on the training set, and their performance is evaluated on the unseen validation set. The model that performs best on the validation set is chosen.
  + **AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion):** These are statistical measures that consider both goodness-of-fit and model complexity. The model with the lowest AIC or BIC score is often preferred.

**Choosing the Right Criteria for Final Model Selection:**

The choice of criteria for selecting the final model depends on your specific goals and priorities. Here's a general guideline:

* If interpretability and avoiding overfitting are crucial, using a metric like adjusted R-squared or a smaller number of predictors might be preferable.
* If the main focus is on achieving the best possible prediction accuracy, using validation set performance or a metric like AIC/BIC might be better.

**Overall, the best subset selection algorithm provides a systematic approach to identify a good model with an appropriate number of predictors from a pool of candidate models.** It's important to consider the trade-off between model complexity and generalizability when choosing the final model.

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

This famous quote by statistician George Box captures the essence of using models in science and data analysis. It can be interpreted in two parts:

**1. All Models Are Wrong:**

* Models are simplifications of reality. They cannot perfectly capture all the complexities and nuances of the system or phenomenon they represent.
* Models often rely on assumptions that may not hold true in every situation.
* As a result, all models will have some level of error or inaccuracy in their predictions.

**2. Some Are Useful:**

* Despite their imperfections, models can still be valuable tools.
* A "useful" model is one that provides valuable insights into the system or phenomenon being studied, even if it's not entirely accurate.
* The usefulness of a model can be judged based on its ability to:
  + **Make accurate predictions:** Even if not perfect, the model's predictions should be reasonably close to real-world observations.
  + **Improve understanding:** The model can help us understand the relationships between variables and identify key factors that influence the system.
  + **Support decision-making:** By providing insights and predictions, models can inform decisions and guide actions even if they don't offer absolute certainty.

**Here's an analogy:**

Imagine a map. A map is a model of a real geographical location. It simplifies the complexities of the landscape, leaving out details like every tree or building. However, a good map is still a useful tool for navigation because it provides a general picture of the area and helps you get from point A to point B.

**Key Points:**

* Don't expect models to be perfect representations of reality.
* Focus on the **usefulness** of a model based on its ability to provide insights and predictions.
* Continuously evaluate and refine models as new data and understanding become available.

By acknowledging the limitations of models while appreciating their potential benefits, we can leverage them effectively to explore data, understand complex systems, and make informed decisions.

1. Self-evaluation
2. Challenges you had during work and how you handled them.

Understanding Data Availability: The initial prompt about downloading data with specific variable names ("Antal personbilar") required me to make assumptions about the data structure available through the PX-WEB API. I highlighted this by mentioning the need to check the API documentation for the exact variable codes.

Error Handling in R Code: You provided code snippets with potential errors (e.g., incorrect arguments in as.data.frame). I analyzed the error messages and suggested appropriate solutions based on the pxweb package documentation and best practices for data conversion in R.

Interpreting Limited Information: When you mentioned high RMSE values, I had limited access to the specific data and model details. However, I focused on general interpretations of RMSE and suggested improvements applicable to various regression modeling scenarios.

Tailoring Responses: I aimed to balance providing comprehensive explanations with keeping the responses concise and focused on the task at hand.

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

Appendix A

Source list

* [**https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/**](https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/)
* [**https://ropengov.github.io/pxweb/articles/pxweb.html**](https://ropengov.github.io/pxweb/articles/pxweb.html)
* [**https://github.com/rOpenGov/pxweb?tab=readme-ov-file**](https://github.com/rOpenGov/pxweb?tab=readme-ov-file)
* [**https://www.scb.se/vara-tjanster/oppna-data/api-for-statistikdatabasen/**](https://www.scb.se/vara-tjanster/oppna-data/api-for-statistikdatabasen/)

**Additional Resources:**