

**COURSE 18-785: DATA, INFERENCE & APPLIED MACHINE LEARNING**

**ASSIGNMENT 4**

**Nchofon Tagha Ghogomu**

**ntaghagh**

*October 14, 2024*

**LIBRARIES:**

The following libraries were used:

* Pandas[1]
* Numpy[2]
* Pyplot from Matplotlib [3]
* Stats from Scipy
* Seaborn
* Linear model from Sklearn
* Linear Regression from Sklearn Linear model

**Programming Language:**

* Python

**INTRODUCTION**

This assignment was made of 4 practical questions that portray real application of data analytics on world data. This assignment was centred around:

* Linear regression with one or multiple explanatory data
* Model fitting and estimation.
* Hypothesis test for correlation.

This assignment had one open study for us to assess and study the trend in the transport domain of transportation.

**SOLUTIONS**

**Question 1: Statistical learning**

* 1. *Describe at least four steps to implementing a rule-based approach to decision-making and give an example. Is any domain knowledge required to establish a rule? Support your answer with an explanation.*

1. Determine the issue:

To implement a rule-based approach, we need to clearly identify and define the issue. It is essential to detailly outline the problem that needs to be resolve or decided on. For example, to curb the spread of the Marburg virus the medical team in Rwanda needs to quickly and efficiently test individuals for Marburg.

1. Establish rules/test cases:

Based on experience, data, and other sources of credible information, the guideline for identifying this issue is defined. In our example, based on the symptoms demonstrated by bearers of this virus, a list of test cases is identified. This could include checking for fever, temperature, vomiting, and contact with infected patients. This will certainly come from the professional experience of the team.

1. Implement the rules:

Subsequently, these test cases are applied on samples and real-world scenarios. In this case study, the medical team will use the established rules during the screening of people for this disease and the results are noted.

1. Test the rules and evaluate:

Finally, these set of rules is evaluated to determine its effectively and based on the results, the rules can be revisited and upgraded for better results. In the case study, the medical team identify the shortcomings of the roles and works upon them.

Domain knowledge is very essential as they guide the team in determining this role. In the above case study, domain knowledge would be understanding the symptoms of the virus and knowing how to test for them.

* 1. *Over-fitting, Simple vs Complex Models.*

This refers to a situation where the model memorises the training data too well but fails to perform on out-sample data. In some cases, the model tends to fit noise, making it overly complex and less reliable.

From the standpoint of parsimony, I would choose the simple model with a single parameter. Simpler models are easily understood while more complex models would make it difficult to see how each parameter contributes to the model.

* 1. *Two commonly used approaches to avoid over-fitting.*

1. Regularisation: In this technique we prevent overfitting by penalising the model using a cost function to prevent/discourage models that tend to be too complex.
2. Cross-Validation: In this case, the data set is split into two: one part is used for training and the other is used to evaluate the model’s performance.
   1. *Provide two examples of metrics used to evaluate the performance of a model and give a formula for each one. Give two examples of applications and appropriate metrics for each case.*

AIC and BIC

* 1. *Why are benchmarks useful in machine learning and give two examples.*

Benchmarks are crucial in machine learning as they help us evaluated and compare the performance of models to guide our selection. For example, we want to develop a model on a given data set. We can build different models using neural networks, support vector machines, decision tree, to list a few. At this stage we evaluate the models using the same evaluation technique as the R-Squared value, F1-Score, MSE, to list a few. Now we perform a benchmark and the best performing model is chosen.

**2. Machine Learning (25 points)**

*2.1 What is machine learning? Discuss its evolution over time and why is it popular?*

Machine learning is a knowledge discovery process based on data and experience. The growing popularity of machine learning is backed by the fact that algorithms are now capable of understanding partens in complex set of data, that would otherwise be very difficult for humans. The daily discovery special use cases, and its application in diverse fields in practical ways have also contributed to its fame.

The concept of machine learning all started when Alain Turing brought in the idea of learning machines in 1950. Later in 1952, Author Samuel of IBM developed ELIZA, the first game-playing algorithm that prepared people for victory against a world champion. In 1957 The idea of neural network was introduced by Frank Rosenblatt with the discovery of the perceptron, a simple linear classifier. In the 1990, the concept of Artificial Intelligence came around as knowledge in computer science and statics were brough together to create different data-driven machine learning approaches. Then there was a rise in Big Data following a boom in volume, velocity and variety of data that could be used for research. There has been improvement in infrastructures, network and standard, under the umbrella of Open Data and IoT, that makes data more and more accessible for use.

In the last decade, there has been growing discoveries in this domain like Google’s Alpha Go, Open AI’s Cha GPT, DeepMind’s Alpha Fold, to list a few. Newer machine learning approaches to solving complex problems are also being developed.

*2.2 Give three examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches.*

Classification: This is a supervised machine learning technique that predicts the category or label of a sample input. To arrive at this stage, the model is trained with labelled or classified data.

Clustering: This is an unsupervised learning technique where the model groups together data of a given input based on feature similarity. As it is unsupervised, the algorithm is not fed with labelled data.

Linear regression: This is a supervised machine learning technique where and algorithm models the relationship between multiple values and fits are linear equation to it.

*2.3 What is the difference between classification and regression?*

A classification model would be able to identify to what category a given data belongs to (categorical output) while a regression model would help predict a value based on given input (numerical output). For example, a classification model will give us information as to weather an email is spam or not while a regression model will give us a forecast of the price of a given commodity in each period.

3 examples

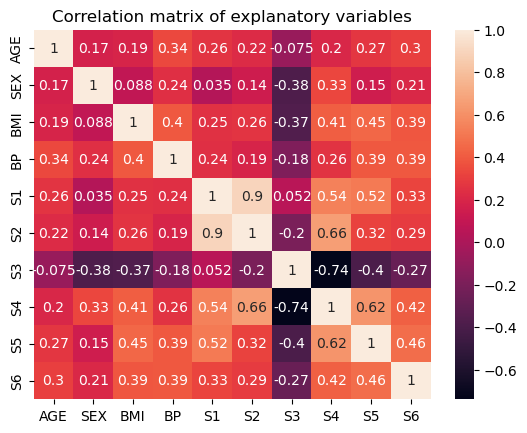
*2.4 What is the difference between supervised learning and unsupervised learning?*

In supervised learning, the model is fed with labelled data in which the desired output is known, and the model learns on it while in unsupervised learning, the model works solely on features. Common supervised learning algorithm include linear regression and neural networks while some unsupervised learning algorithms include the K-means, Principal component Analysis (PCA).

*2.5 Give examples of successful applications of machine learning and explain what technique is appropriate and what type of learning is involved?*

**3. Diabetes data (25 points)**

***3.1 Correlation Metrix***

The correlation matrix shows the correlation between the explanatory variables, with values ranging from -1 to +1. A +1 indicates a high positive correlation, while a -1 indicates a high negative correlation. A 0 indicates an absence of correlation between the variables. Using the pandas *.corr()* function, we observe the following correlation between the variables.

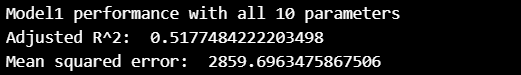
* We observe the greatest positive correlation between the blood serum measurements S1 and S2, and the greatest negative correlation is between the blood serum measurements S3 and S4.
* On the wider scale, S3 demonstrates negative correlation with other serum measurements.

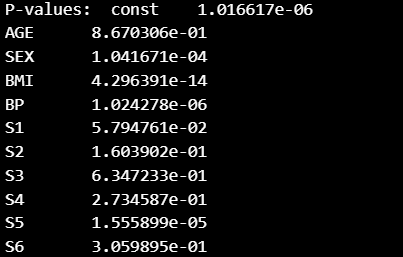
***3.2 Coliniarity***

Collinearity between predictors refers to a situation where the predictors are highly correlated. This can be identified when the calculated correlation between them tend to 1. In this exercise, we identify S1 and S2 as colinear features since their correlation, 0.9, is very close to 1.

Collinearity between features can cause high instability in their estimated coefficient value. This makes it highly sensitive to very small variations and can cause the model not to be generalisable.

***3.3 Model 1***

Using the linear regression from the Sklearn linear model, Model1 is built. We obtain the following MSE and R-squared for Model1 as seen bellow. The and R-squared is acceptable as very high values could be indications of overfitting.

****To analyse the significance of every feature, can compute the p-values and values bellow the set 0.05 value, we will consider them as insignificant. Bellow are the p-values of all features.

From these results, we notice that not all features are significant. Also, this can partly be due to collinearity. Of features that demonstrate high collinearity, one of them can be drop since their variation is already represented by the collinear variable.

***3.4 Forward Vs Backward Selection***

These are methods use in selecting features for a model. In forward selection, we begin with an empty list of variables. Any feature which significantly adds to the overall R-squared value of a model is added to the list of variables, starting from the most significant. Any insignificant feature is left out. On the other hand, with backward selection, stepwise, we remove variables that don’t significantly contribute to the overall performance of the model till the point. These selection criteria yield different set of variables as variables tend to influence each other when in a model.

***3.5 The Stepwise Approach***

In the approach, a variable added or removed from the list of selected features is subject to scrutiny and can be removed or added respectively, steps down the process. For example, if a feature is added because its large individual power, its performance is evaluated again in the model and if its contribution is insignificant, it can be removed.

Using stepwise forward regression, we obtained the following features:

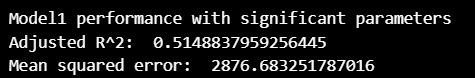


The stepwise function works as such:

* In stepwise forward regression, the variable based on their individual strength are added one by one to the list, starting with that of greatest strength.
* At each step, all variables are evaluated on their individual contribution to the overall model.
* Feature that does not significantly contribute to the overall model are removed and the cycle continuous.

Basically, a feature can appear significant but when added to the model with other features, its contribution might be minute. In this case, the feature is dropped.

The Mean Squared Error and R-squared value for this new model was found to be:



The R-squared value for the new model is almost the same as that of the model with all parameters. It is lower than Model1 by 0.0029. The MSE is of this new model slightly higher than that of Model one by a factor of 16. On a broader scale, both models are similar.

*Conclusion*

The model based on the best features from stepwise regression is roughly as productive as the model will all 10 parameters. However, this model is more parsimonious as it has fewer parameters but equally as productive.

**4. Analyzing the Titanic data set (25 points)**

***4.1 Logistic Versus Linear Regression***

They are both methods of supervised learning but differ in that logistic regression is used to predict binary outcomes while linear regression is used to predict continuous outcome. In this example, we want to predict survival or not based on the parameters. Other examples and application of logistic regression is in healthcare when we want to know weather or not a drug is effective. Some example scenarios when linear regression is applied includes when we are trying to predict stock prices or what GPA a student will have given the parameters.

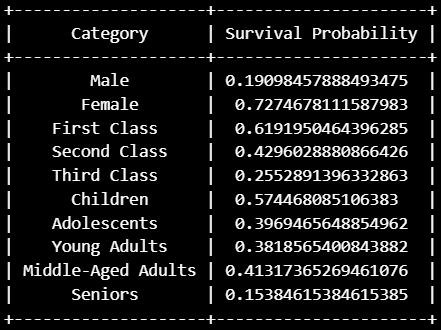
***4.2 Probability of Survival of a Passenger***

To calculate the survival probability of a passenger:

* The passengers were grouped by gender, passenger class, and age group.
* For each group, passengers were broken down into Male, Female for gender, 1st, 2nd and 3rd class for passenger class, and Children (0 - 12years), Adolescents (13 – 19years), Young Adults (20 – 34years), Middle Aged Adults (35 – 64) and Seniors (65 +)
* For each category, the probability of survival was calculated by dividing the number of survivors by the total number of people in each category. In a given category:

***4.3 Probability table***

The following result was obtained:



It is observed that being a female, one had the greatest chance of survival followed by being in first class, then being a child. The least probability of survival was in the Male category. On can obtain the combined probability by multiplying categories together.

***4.3 Logistic Regression Model***

**Question 1:**

1. Goal:

This question is based on linear regression with one explanatory variable. Here, we build a model using FTSE100 monthly return, which will be our dependent variable, and house prices monthly returns as explanatory variable. In the end, we conclude on the result and the correlation coefficient observed.

1. Steps:

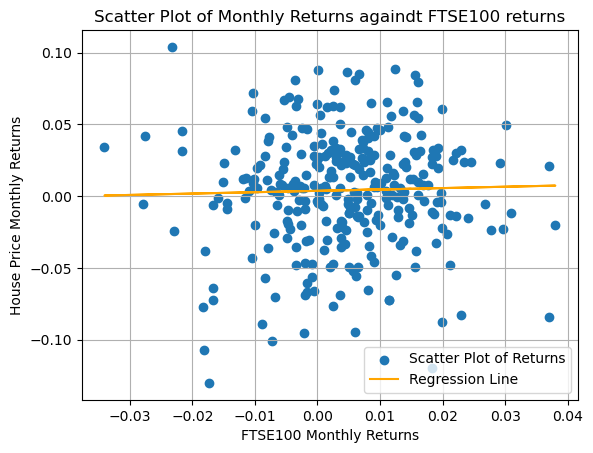
To arrive at the plots and insights, the following steps were used:

* The required data was imported.
* Fields, like the “Date”, was renamed and converted to the appropriate date format.
* Monthly return for both data sets are computed using the *pct\_change()* function.
* A regression model is created using *linregress* of the imported *stats* module.
* Using the intercept and slope from the previous step, we can plot our regression line.
* A scatter plot is plotted.
* The obtained gradient and correlation coefficient of the line is observed and commented on.
* We use the Pearson correlation to measure the linear relation between both variables.

1. Results and observation:

Here, we explore the broken-down result of this exercise:

1. The regression model:

Below is a graph showing our model and a plot FTSE100 returns against average house prices return. The slope and intercepts were respectively 0.0955 and 0.0036

A number on a black background

Description automatically generated

1. Observation:

There is very little or no correlation between both variables. The slop is very gentle and an increase in FTSE100 is not reflected by a significant increase in How

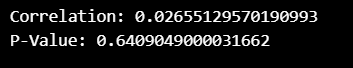
1. Hypothesis test:

The null hypothesis, μ0:

* There is no correlation between the FTSE100 and the average monthly return.

Alternative hypothesis μ1:

* There is a positive correlation between the returns of FTSE100 and the monthly return.

In this experiment we will be using a one tail test to confirm positive correlation. An α = 0.05 is used to perform this test and we observed a p-value of 0.64

Drawn from this result, the null hypothesis is accepted because p = 0.64 is way greater than α = 0.05. Therefore, we accept the null hypothesis. We can safely conclude that there is no significant correlation between FTSE100 and House Prices. Some explanation to this can be:

* FTSE100 and House Prices belong to different asset classes: equity and real estate respectively. They are both
* Triggers. The dynamics of FTSE100 and House Prices are influence by largely different variable. For example. FTSE100 can be affected by corporate earnings, investor sentiment, and macroeconomic indicators while House prices can be influenced by the socioeconomic conditions and demand.

Though the overall state of the economy can indirectly influence both, as can be seen in the very low correlation, they are both largely uncorrelated.

**Question 2:**

1. Goal:

Perform linear regression on multiple explanatory variables drawn from a typical academic dataset, to estimate the graduation rate of a school.

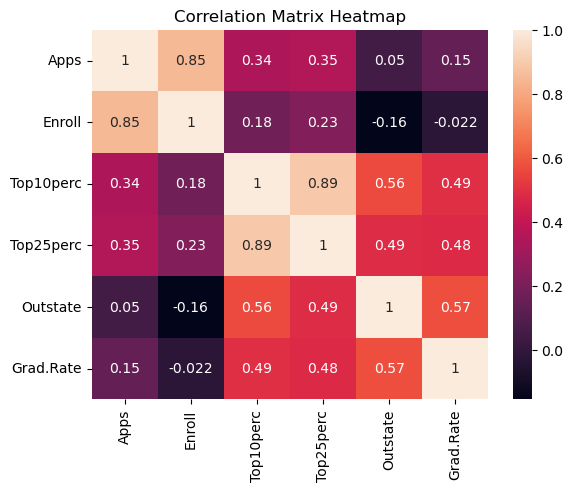
1. Steps:

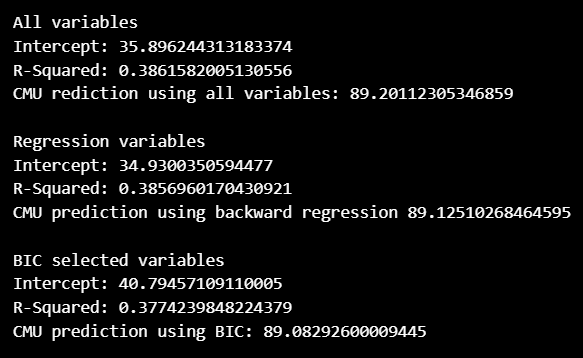
To arrive at the results and insights:

* The required data was imported as variables.
* The correlation metrics is calculated using panda’s *.cor()* function.
* Stepwise backward regression is performed to select features with alpha = 0.5
* Predictor variables are identified with explanations.
* Using BIC, predictor variables are selected.
* The accuracy of the chosen models is calculated and compared with the five predictor model
* The result is annualised.

1. Results and observation:

Bellow are the results and observations achieved.

1. The correlation between the variables can be represented using the heat map bellow:
2. A stepwise backward regression was performed to select predictors. In this process, we commence will all independent variables, and by fitting the model to each feature, we eliminate the least significant values using the p-value threshold of 0.05, till all features are statistically significant. To ease this process, we used a function proposed by Askkash[4].
3. Using backward regression, we fall on the following predictor variables: Apps, Enroll, Top25perc, and Outstate
4. Using the BIC model, we obtain a set of different parameters: Apps, Enroll, Topperc, and Outstate. Just 3 of these variables are common and in the place of Top25per for the stepwise regression, we have Top10perc for BIC. But their accuracies are very similar which we will se in the next point.
5. Comparing the accuracies using their R-squared, value we noticed that for all three models, (all five, backward and BIC), their accuracies are very much similar. Using all the variables, our model is the most accurate with an R2 score of 0.386 but just different from BIC by a factor of 0.01. BIC model can be used in this case if we are considering parsimony.



1. The most accurate model is the one that uses all five features and by using this model, we obtain a graduation rate of 89.2 as seen above.

In conclusion, we see that using all the features in this case is the most accurate and defers very slightly from the BIC model. This implies that using fewer models can still yield almost the same result. If considering computation power optimisation, using fewer features will totally work out.

**Question 3:**

1. Goal:

Undertake an open study to assess the trend in domain of transport in any country or group of countries of my choices. This should be based on publicly available data.

1. Steps:

The data used for this open study all come from the World Bank Indicators and are:

1. Mortality caused by road traffic injury (per 100,000 population)
2. Number of deaths ages 20-24 years

Sub-Saharan African Countries were used for this study. We will be analysing the relationship between death in young adults and the death caused by traffic injuries per 100,000 population. The steps used are:

* Both data frames were downloaded and imported from the World Bank Indicator.
* Data from selected counties/region is extracted
* The null and alternative hypothesis test is identified
* A scatter plot of both indicators is plotted.
* A linear regression model in developed with Mortality caused by road traffic injury as independent variable and number of deaths ages 20-24 years as explanatory
* The model is plotted
* The correlation coefficient and p value are calculated, and the hypothesis is tested.
* Predict the value for 2021 and conclude.

1. Results and observation:

We arrived at the following:

1. Hypothesis testing

The null hypothesis, μ0:

* There is no correlation between deaths of 20 – 24 years and the Mortality caused by road traffic injury (per 100,000 population)

Alternative hypothesis μ1:

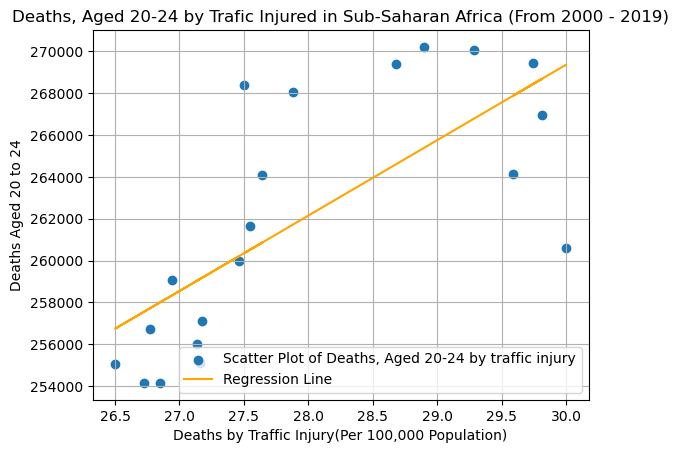
* There is a positive correlation between both variables.

A black background with white numbers

Description automatically generatedWith α = 0.05 we perform a Pearson test of correlation the following p-value was observed. With a P value of 0.00038 << α = 0.5, we conclude that there is a significant relationship between both values. For emphasis, a one tailed test is performed since we want to check for a positive correlation only.

1. Plots:

The following scatterplot of both variable and line plot of the model was obtained. We observe a strong correlation between both variables as increased in death by traffic injury per 100,000 is reflected in an increase in deaths of 20 – 24-year-olds, with a slope of 3606, and an intercept of 161170.



This makes sense as youths of this age group are prone to risky behaviours including reckless driving, and they also spend a deal of time on the road: commuting to school, work to name a few. By knowing this, regulations can be made to make the road safer for this vulnerable age group and sensitised them about these findings.

1. Predictions for 2021:

Give there is no data for the total number of deaths from road injury for 2021, based on persistent forecasting, the best guess will be the last value with (2019) which is 27.4599 deaths per 100, 000 population. Using this, we estimate the number of deaths of 20 – 24-year-old at 260199 deaths.

**Question 4:**

1. Goal:

Asses the Israeli bank data on employment and predict the unemployment rate for 2020, explain to compute accuracy estimates and present the estimates in the form of percentages.

1. Steps:

To arrive at the plots and insights, the following steps were used:

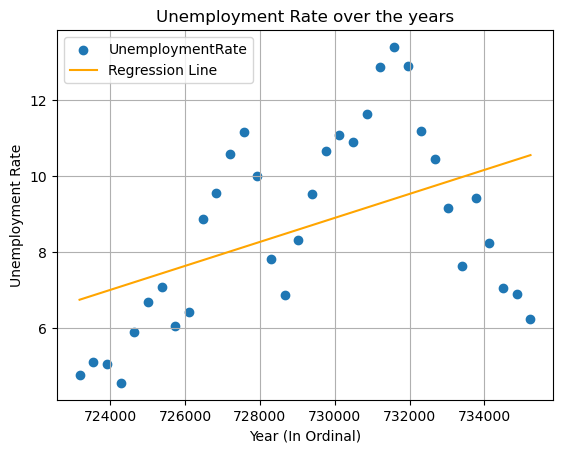
* Dataset is downloaded and extracted into data frame.
* Data from 1980 to 2013 is extracted and the date is converted to ordinal using the toordinal() python function.
* A linear regression model is fitted using date as independent variable and unemployment as explanatory variable.
* We predict the unemployment rate for 2020
* The accuracy is evaluated using MAPE
* Conclusions are drawn.

1. Results and observation:

We observed the following.

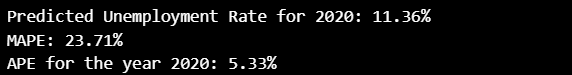
1. Model and prediction

Using linear regression with the LinearRegression() model, we fit the date in ordinal form as independent variable and unemployment rate as explanatory variables. We convert the year 2020 to ordinal and fit it in our model to estimate the values of unemployment. The following result was obtained for the likely rate of unemployment.

To have a better appraisal of trend of unemployment and the model, the following graph was plotted.

1. Accuracy evaluation:

There exist a host of accuracy evaluation method and they all explain in different ways how far the mode is from the actual value. In this case, we us the Mean Absolute Percentage Error (MAPE), which is a method of evaluating accuracy by expressing the ratio error to the actual value, in percentage form, and obtaining the mean of individual averages. This can be expressed using the formular bellow where Ai is the actual value, Fi the forecasted value and n the number of observations.

We obtain an MAPE value of 23.71% and the average percentage error for the year 2020 is 5.33%. This value suggests that on average, the values are 23.7% off the actual value. Based on the analyst’s threshold expectation, this model can be analysed as suitable or note.

**REFERENCES**

[1] “pandas documentation — pandas 2.2.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/index.html

[2] “NumPy -.” Accessed: Sep. 02, 2024. [Online]. Available: https://numpy.org/

[3] “Matplotlib documentation — Matplotlib 3.9.2 documentation.” Accessed: Sep. 02, 2024. [Online]. Available: https://matplotlib.org/stable/

[4] A. R. V, *AakkashVijayakumar/stepwise-regression*. (Mar. 27, 2024). Python. Accessed: Oct. 15, 2024. [Online]. Available: https://github.com/AakkashVijayakumar/stepwise-regression