DATA, INFERENCE AND APPLIED MACHINE LEARNING

ASSIGNMENT 4

CODE : 18-785

CARNRGIE MELLON UNIVERSITY AFRICA

ANGE IZABAYO

MS ECE'25

Email: aizabayo@andrew.cmu.edu

Submission date on October 23, 2023

REPORT

Introduction

For assignment 4, I utilized Jupyter notebook for coding. The assignment comprised of 4 questions mostly focus on performing various form of linear regression in different forms. In the first cell, I initialized the libraries necessary to fulfil the assignment. Means all libraries I utilized were in the first cell as I am performing, I kept updating them. I utilized: pandas which is used in manipulation and structuring of the dataset. I used NumPy library which used in working with array which is mostly where I used it and also to perform mathematical function. I used also SciPy which also python library, which is mostly used in statistical analysis, testing and visualize data. I used scikit-learn library known as sklearn which is mostly used in predictive data analysis.[1] sklearn library provide various tools for modelling and machine learning including regression, coefficient of determination r square, for mean absolute error, linear model. I used also statsmodels which is also python library which help in estimation of different statistical model and hypothesis testing. Additionally, I used quandl to access the only data of Israel unemployment, where it provides opportunity of accessing data online or download them.

These libraries, along with other built-in functions, were used to tackle various task provided.

Q1.

In the first question the task is to use two dataset one of Monthly house price data in UK and FTSE100 index to create regression model, calculating correlation coefficients and performing hypotheses test to back the result obtained.

I began by using pandas to load both dataset in Jupyter notebook as dataframe, then I extracted the necessary columns to help me performing.

Since the dataframe of house price has data column with date beyond the required on I removed unnecessary one and then set date column as index by using python built-in function set index () and due to the date in FTSE100 dataframe were not sorted in ascending order I sorted the date and format the date time to match with one of house price.

I proceeded by finding monthly return for both dataframe I used built-in function percentage change(pct_change) and there is also mathematical formula to find it. after finding the monthly return I used dropna () function for removing the rows that contain NaN values.

After finding and cleaning the monthly return on both data frame I declared the variable x containing column of house price monthly return and y with Adj close monthly return to be used in scatter plotting and finding the regression line.

I used matplotlib library which is used in data plotting and visualization using graph, I made scatter plot of dependent variable FTSE 100 return of against house price monthly return as independent variable.

By using SciPy library I employed linregress () to predict the continuous relationship between dependent and independent variables (Correlation coefficient), to find the slope, p_value which measure the significance relationship of two value, and also to estimate intercept, standard error to measure the uncertainty in estimated slope. Then i performed mathematical operation of multiplying slope to independent variable and then add intercept to find the regression line. I used matplotlib to plot to the same graph with the scatter.

Used if statement to show whether there is significant relationship between dependent variable and independent variables or not, by taking p (P_value) from the SciPy calculate lingress () function.

I also extracted correlation coefficient from that SciPy function

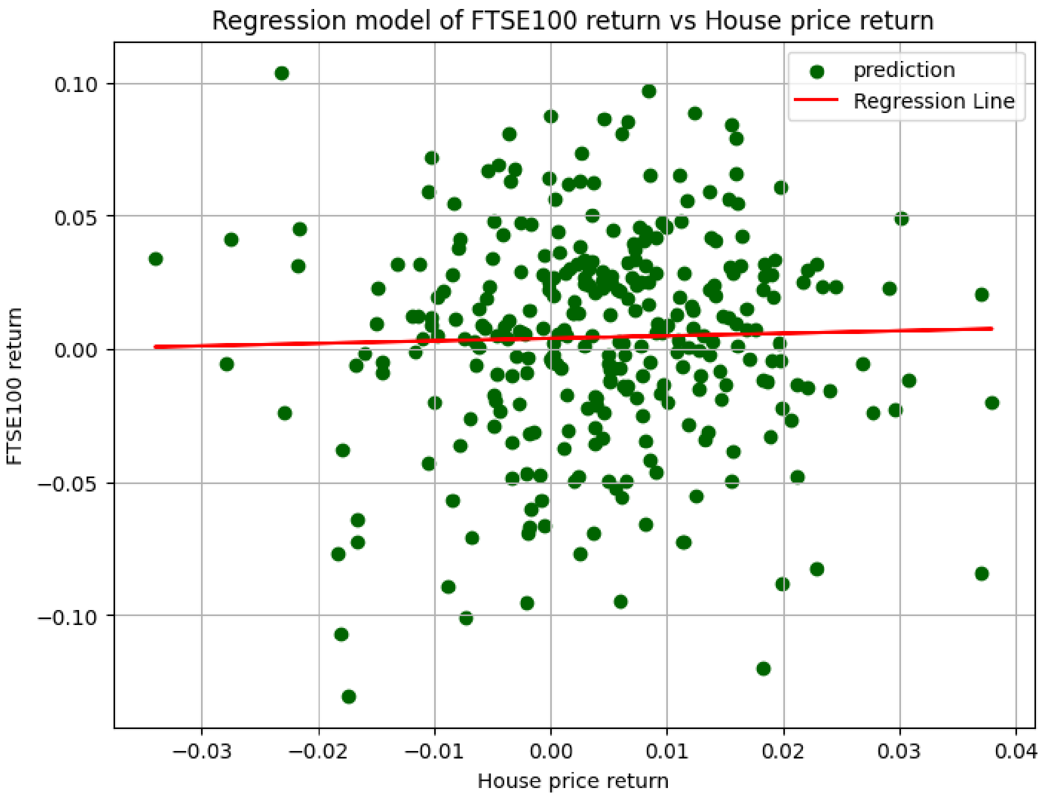
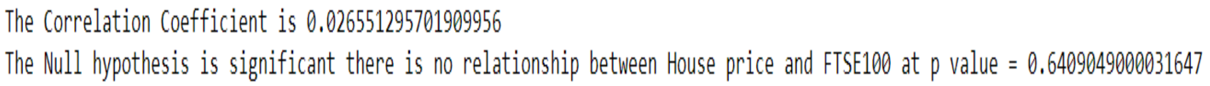


Figure 1Monthly return of FTSE 100 against Average house price and regression model



```
The Correlation Coefficient is 0.026551295701909956
The Null hypothesis is significant there is no relationship between House price and FTSE100 at p value = 0.6409049000031647
```

Based on the result the obtained Coefficient of correlation is 0.0265 which is positive means that FTSE100 monthly moves in the same direction as Average House price monthly. But this correlation is very small which is approximately to 0 shows that there is no relationship between them, but in this context the correlation I obtained if I consider all decimals indicate the increase in one variable and other variable increases too but on low rate.

By using P value with significance value equal to 0.05 the p value is 0.64 which prove that there is no linear relationship between house price return and Ftse100 return.

In conclusion since the correlation coefficient shows very small value indicating that the increase in Ftse100 tend to increase with increase in average house price and the null hypothesis, which is significant as value close to 0.05, I can draw conclusion that there is almost no relation between two index.

Q2.

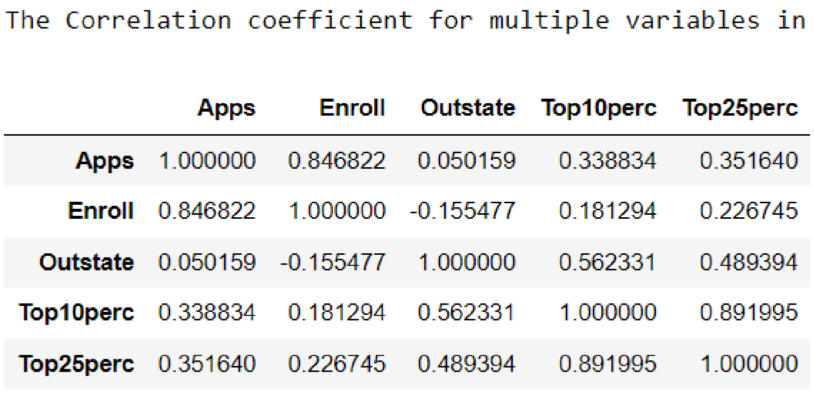
The question 2, given dataset of containing information of US college and universities, where we are supposed to use few column to perform the task include: Application column (Apps), enrolled student (Enroll), out of state student (Outstate), admitted top10% (Top10perc), and admitted top25% (Top25perc). These are the column or independent variable to be used to predict graduation rate. Performing linear regression with multiple variables.

I started by reading the dataset in Jupyter notebook contain all information using pandas library.

A) Task a is to find the correlation coefficient for multiple variable

The approach to the task I first extracted the column to be used from the dataframe and then store them in declared in variable. By using pandas function corr() I determined the requested correlation coefficient of the dataset.

Table 1 Correlation Coefficient



```
The Correlation coefficient for multiple variables in College dataset
```

	Apps	Enroll	Outstate	Top10perc	Top25perc
Apps	1.000000	0.846822	0.050159	0.338834	0.351640
Enroll	0.846822	1.000000	-0.155477	0.181294	0.226745
Outstate	0.050159	-0.155477	1.000000	0.562331	0.489394
Top10perc	0.338834	0.181294	0.562331	1.000000	0.891995
Top25perc	0.351640	0.226745	0.489394	0.891995	1.000000

The table 1 shows the output of Correlation coefficient calculated this table shows the correlation coefficient for each column compare to other to find their relation and return the result in form of 5x5 matrix and most of the relation shows are positive which indicate the variable tend to increase with increase in other in other world they move in the same direction except for combination of Enroll with Outstate and for Outstate with Enroll shows negative correlation which means moves in opposite direction.

B) Asked to consider Graduation rate as dependent variable and use stepwise to build linear regression model.

I began this task by declaring variable to store the data extracted from dataframe. Where one variable contains independent variables which are five and dependent variable containing graduation rate.

I defined function forward regression to perform stepwise regression and I have initiated the parameter independent variable to store predictor here are Apps, Enroll, Outstate, Top10perc, Top25perc. And dependent variable which is Graduation rate, significant level (Alpha), and Verbose which is set to False by default, this is flag which help in understanding the process and monitoring execution

In this code I need to predict the best fit in different features.

I started by creating an empty list that will store the best fit, and after creating the loop where whole process will be performed. Proceed by creating variable that will keep store the column which do not fit using for loop to iterate in the feature and the one which is not selected stored in eliminated variable.

By using significant level alpha, which is equal to 0.05, by using if statement to set p_value to be less than alpha and after appending the best fit feature to the declared variable.

By using the imported statsmodels.api library I declared in the first cell I employed sm.OLS function to perform linear regression model and used fitting function to fit the model to find the standard error, p value, r squared. To estimate the best features, I will be using the p value if the p value of the feature is less than the significant value or threshold the feature will be stored in predictor initiated and if is p value is greater than alpha it will be stored in other variable in predicting the graduation rate.[2]

I have printed the best features and also summary containing regression analysisBy using matplotlib library I have created the scatter plot with regression line of the best fit

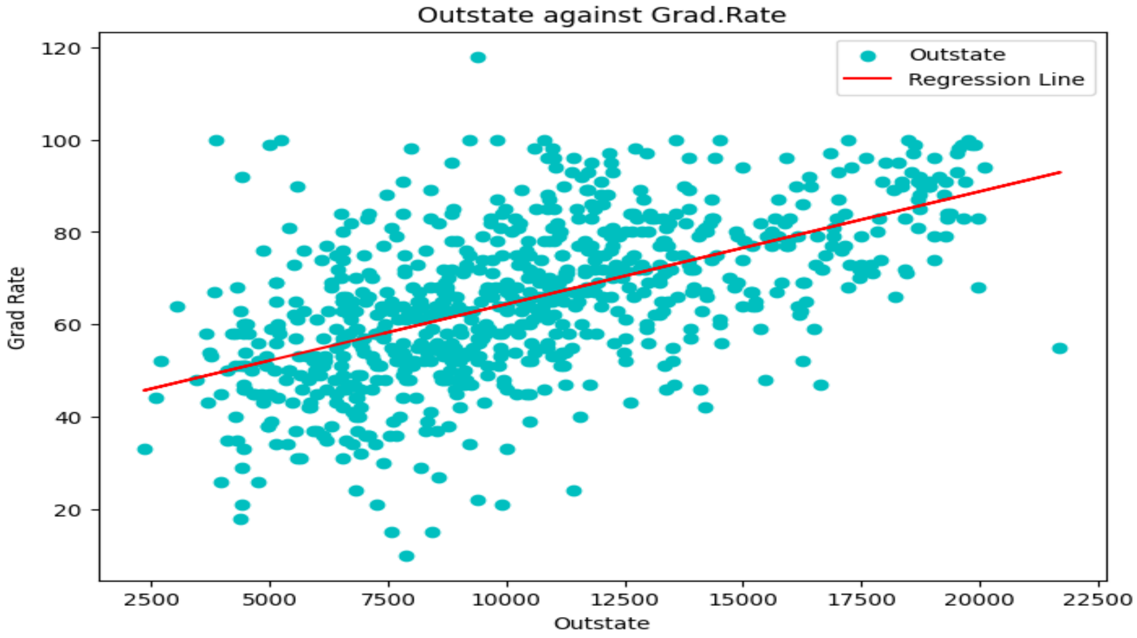


Figure 2 Outstate against grad rate

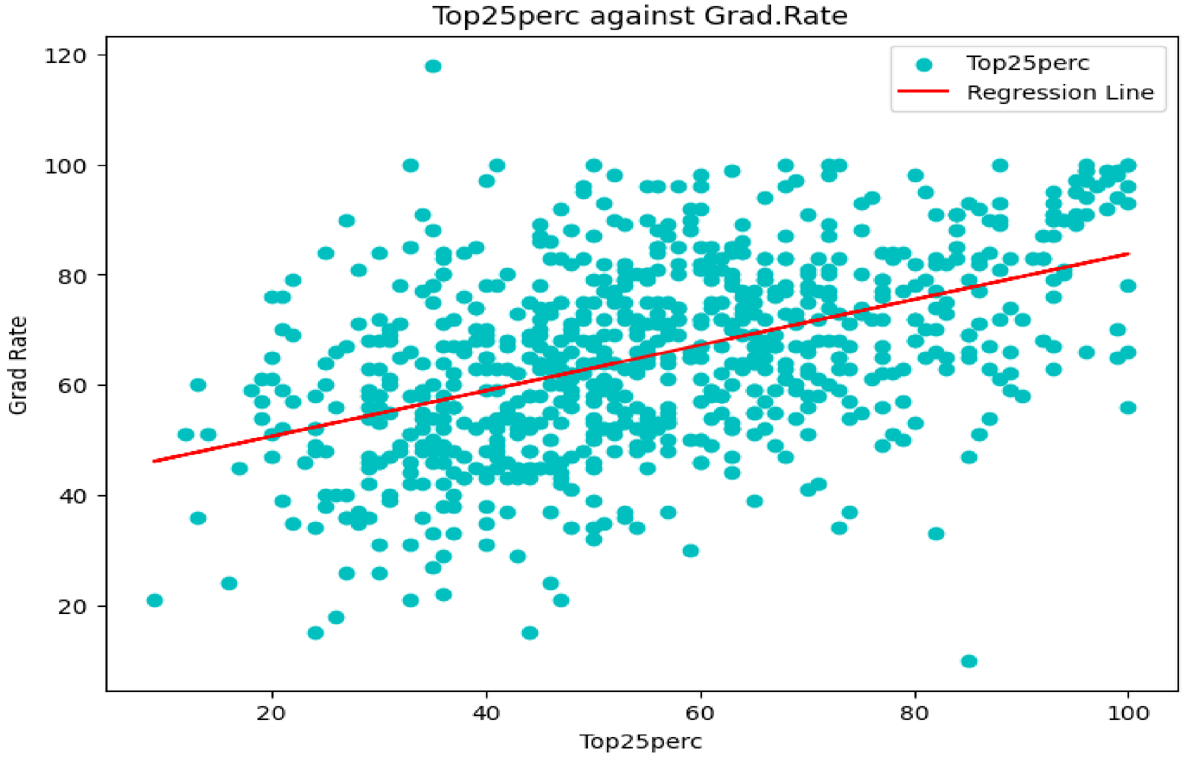
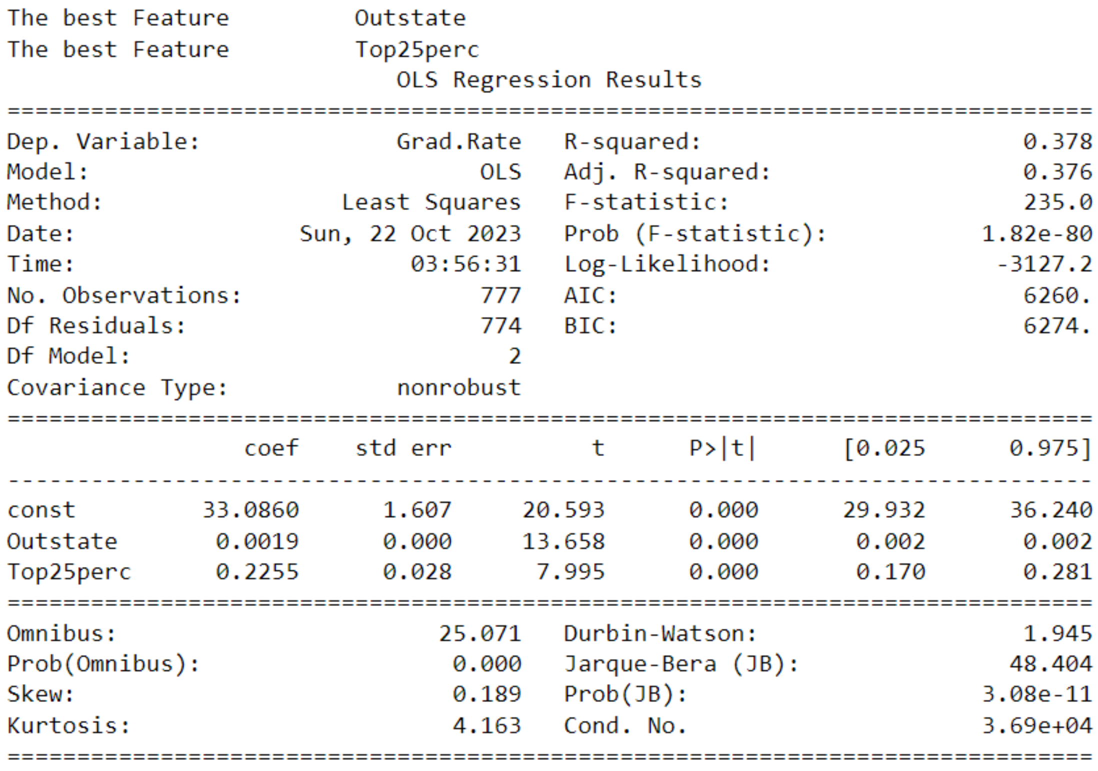


Figure 3 Outstate vs Grad Rate



```
The best Feature          Outstate
The best Feature          Top25perc
                            OLS Regression Results
==============================================================================
Dep. Variable:              Grad.Rate   R-squared:                       0.378
Model:                            OLS   Adj. R-squared:                  0.376
Method:                 Least Squares   F-statistic:                     235.0
Date:                Sun, 22 Oct 2023   Prob (F-statistic):           1.82e-80
Time:                        03:56:31   Log-Likelihood:                -3127.2
No. Observations:                 777   AIC:                             6260.
Df Residuals:                     774   BIC:                             6274.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.0860      1.607     20.593      0.000      29.932      36.240
Outstate       0.0019      0.000     13.658      0.000       0.002       0.002
Top25perc      0.2255      0.028      7.995      0.000       0.170       0.281
==============================================================================
Omnibus:                       25.071   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.404
Skew:                           0.189   Prob(JB):                     3.08e-11
Kurtosis:                       4.163   Cond. No.                     3.69e+04
==============================================================================
```

Figure 4 Summary

From the figure 4 shows the best features which are Outstate and Top25perc and the summary of regression model. This are the best predictor variable useful in predicting the graduation rate. Where I have used regression model analysis and p value is the one used to determine the best fit by comparing to the threshold, where p value must be less than significant level/threshold (0.05) to be chosen

In conclusion I have learnt stepwise regression how to perform model and fitting on multiple variables to determine the best fit.

C) Asked the useful predicator variable in predicting the graduation rate.

The best predicator variable as described in sub question B are Outstate and Top25perc. This variables were selected based on the condition I have set in the forward regression function, whenever the p value obtaining in the fitting the regression model is less than threshold which is set to 0.05 this will be selected as the best fit and if the feature is above the threshold, it will be eliminated.

D) Tasked to evaluate the usefulness of predictor variables in predicting graduation rate by using BIC (Bayesian Information Criterion.

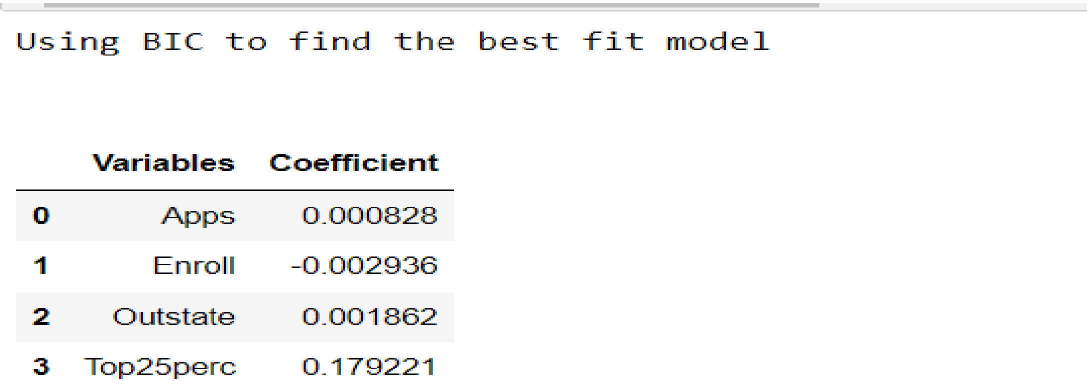
Bayesian Information Criteria is another statistical method that balances the goodness of fit and model complexity. It helps in model selection by considering both how well a model explains the data and the parameters used, and it help in selecting the best appropriate model.[3]

For this task I have to use five features which are predictors to predict the graduation rate using this method.

Process I have used lassolarsic function which is sklearn library with the BIC criterion to automatically select the most relevant features for linear regression model. And the use the Model to fit. Where it takes independent variable (predictor) and dependent variable graduation rate to select the best features based on BIC criterion. The selected features must have correlation significantly to the Graduation rate.

The selected features or the fit variables has coefficient which determine the strength and the direction of the featured in relation to the dependent variable (Grad. Rate)

The I have created dataframe for better presenting the selected features with their corresponding coefficients.



```
Using BIC to find the best fit model
```

	Variables	Coefficient
0	Apps	0.000828
1	Enroll	-0.002936
2	Outstate	0.001862
3	Top25perc	0.179221

Figure 5 The BIC Best fit variables

The result displayed in figure 5 confirm the features selected using BIC method. Notably, these features include Apps, Enroll, Outstate, and Top25perc. The coefficient associated with each feature indicate the relationship between independent variables to graduation rate. Where Apps, Outstate, and Top25perc has positive regression coefficient, meaning that the feature has the same direction with the graduation rate, whenever these feature increases the graduation rate increases. While for Enroll has negative Regression coefficient which indicate that there is negative relationship. As the number of student enrolment increases the graduation rate decreases.

BIC method is simple method compared to stepwise regression for statistical modelling and feature selected where it has increased the feature compared to the previous method stepwise method I have used.

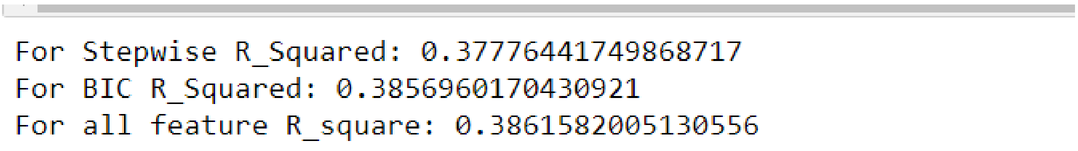
E) Comparing the accuracy of the model using BIC model and Stepwise model using only useful variable, and then find the accuracy of the model using all five predictors.

In response to this task, I calculate the accuracy of the selected features in Stepwise model and BIC model using R-squared(r2) metric. R-squared metric is a valuable measure when assessing the regression model, for here I have employed it for predicting the graduation rate. Where it quantifies the proportion of the variance in the target variable that can be explained using selected features. The higher R-squared value indicate the better accuracy, signifying that the chosen features are more effective at predicting variations in graduation rate. I will be comparing the accuracy of the predicted features from Stepwise model to the one of BIC model[4].

Staring from stepwise model to determine the accuracy of selected features I have used statsmodels.api library which I have used in stepwise modelling to predict the goodness fit and then used r2_score the sklearn function used in calculating the R-square

For the Bayesian Information Criteria (BIC) model, I have started by creating the dataframe to store the selected features found using BIC model, and then configure the model to fit the relationship between the selected features which are Apps, Enroll, Outstate, Top25perc and dependent variable, which is graduation rate. And then predict the graduation rate using predicted values. By using r2_score () function of sklearn library to calculate the R-squared between predict variable graduation rate and predicted values, and then print the result.

Calculating the accuracy of the model when I employ all five predictor to predict the graduation rate. I have used linear regression model for all predictor to predict the graduation rate. I created the instance of linear regression model. And then train the model using all feature include Apps, Enroll, Outstate, Top10perc, and Top25perc. With dependent variable representing graduation rate. Predict the Graduation rate based on trained model features. After I used the r2_score function to find the R-squared between prediction generated by the model.



```
For Stepwise R_Squared: 0.37776441749868717
For BIC R_Squared: 0.3856960170430921
For all feature R_square: 0.3861582005130556
```

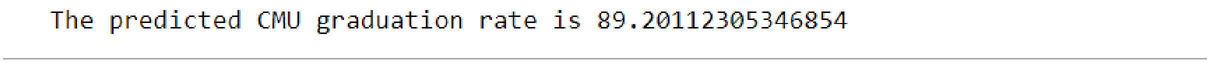
Figure 6 R-squared of different model

Comparing the accuracy of the models that employ useful predictors versus the model using all five predictors reveals interesting insights. Figure 6 displays the R-squared scores, with the stepwise model achieving a score of 0.3777 (37.77%), the BIC model scoring 0.3856 (38.56%) with selected features, and the model utilizing all features reaching 0.3861 (38.61%). These scores signify the predictive power of each model, and higher R-squared values equate to superior accuracy in estimating the graduation rate. Notably, the model that attains the highest accuracy is the one employing all available features, highlighting its potential as the preferred choice for graduation rate estimation. This comparison underscores the significance of feature selection and the added value of including all predictors in enhancing predictive accuracy. And followed by BIC model which has higher accuracy than stepwise regression model.

From this question I have learned linear regression modelling by using different method including stepwise regression, BIC model and the way of testing the accuracy of the model which reflect to what we learnt in class session.

F) Task to predict the graduation rate for Carnegie Mellon University Using the model that has provided the high degree of accuracy.

To approach this question, I started the By extracting data for Carnegie Mellon University in the dataset of colleges and then called the trained model for all predictor, used in the previous question because it is the model that has higher accuracy compare to other method as proved, then predict the graduation rate of Carnegie Mellon university.



```
The predicted CMU graduation rate is 89.20112305346854
```

The estimated graduation rate for Carnegie Mellon University is 89.20.

The estimated Graduation rate of Carnegie Mellon University is fall beyond the actual given in data, which could have caused by error in the model or overfitting.

Q3.

Asked to do study to assess the trend in transport domain using available data and asked to predict the situation in 2021.

Introduction

In this study, I investigated the relationship between Mortality caused by road traffic injury (per 100,000 population) and Death rate, crude (per 1,000 people) in Sweden. In this study I focused on assessing the contribution the contribution of road traffic injury mortality to the overall death rate. The study has objective of determining the correlation between these two variables, analysing historical trends, and predicting the mortality caused by road traffic injury rate for the year 2021 by using mathematical linear regression modelling. By delving into this analysis, I seek to understand and provide insight to road transportation accident dynamic in Sweden.

Data Collection

For data sourcing, I started by searching on World Health Organizations since my study related to health, but unfortunately the data I found was not complete has few years. However, I successfully obtained the dataset with all details from World Bank Indicator, This two dataset I used is of Mortality caused by road traffic injury (per 100,000 population) and other for Death rate, Crude (per 1,000 people) for multiple years ranging between 1960 to2022[5],[6].

Assumption

For this study, the I assumed there is relationship between two variables (Mortality caused by road accident, Death rate) is linear and have no outliers and also all variables are quantitative.

Assuming that by studying the historical trend can be indicative of future trends in Mortality caused by road accident in predicting the rate in 2021.

Methodology

The methodology for this study is in the following step

Data Collection: Gathering data for Mortality caused by road traffic injury (per 100,000 population) and Death rate, Crude (1,000 people) case study of Sweden downloaded fetched dataset from World Bank Indicator.

Data Analysis:

- For this study I will be comparing data for the number of Mortality caused by accident to the death rate in the timeline from 2000 to 2019, which means I will be cleaning data and remove NaN values by ensuring I remain with Quantitative number within the same timeline.
- Understanding the relationship between the two dataset by finding the correlation coefficient.
- For data visualization to understand the trends and relationship I will be using matplotlib python library.

Modelling: I will be fitting linear regression model for variables data to help to help in predicting Mortality rate caused by road accident in 2021 and find its accuracy. To assess how it have changed.

Implementation

To perform the required statistic, I used Jupyter notebook.

To start, I used pandas library to retrieve both dataset and load them in Jupyter notebook, and it contained different countries data. From dataframe I started by extracting the country which is Sweden and cleaning data by removing NaN values from different years where I remained with data from year 2000 to 2019 and column of numbers or values. By using melt function, I have shaped the dataframe individual and rename them to make them suitable for analysis.

By using built-in function merge, I have merged both dataset since their have the same period, I have merged on date and then remain with three column one of date, column of Road accident, Death rate in the country.

I declared predicator variable which represent independent variable and contain the column of date in merged file. I declared predict variable which will be based on data know as dependent variables and it is column in merged dataframe of road accident.

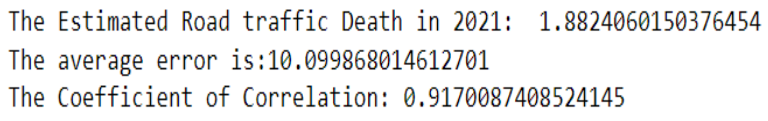
Creating linear regression model by using imported sklearn library I employed linear regression function I declared variable to store the created linear regression model and then after I used the model to fit the data. In fitting I have reshaped the predictor variable in order to be able to determine the best fit of the data to help in making prediction of the mortality caused by road accident in 2021.

By employing multiple library like sklearn, NumPy I trained the model to predict the mortality caused by road accident in 2021 and using NumPy to shape the form the predicted date into an array to allow estimation of expected mortality in 2021.

To find the accuracy of the prediction performance compared to the actual data using Mean Absolute Percentage Error (MAPE) where the lower the average the higher the accuracy and better fit of the model.

I calculated the correlation coefficient by using built-in function corr() to determine the relationship between Mortality caused by road injury and Death rate in Sweden in different period of time.

I plotted two different graph to visualize and explain the relationship or the change in Mortality caused by road injury and Death rate in Sweden. The first graph is scatter plot by Employing matplotlib library I created the graph of Mortality caused by road injury against Death rate. Other graph is timeseries graph where Mortality accident and death rate are against time series. In addition to I have performed hypothesis test to test if the Null hypothesis is significant which implies that there is no relationship between Mortality caused by road accident with overall Death rate in country I used SciPy function stats.lingress to find p value.



```
The Estimated Road traffic Death in 2021:  1.8824060150376454
The average error is:10.099868014612701
The Coefficient of Correlation: 0.9170087408524145
```

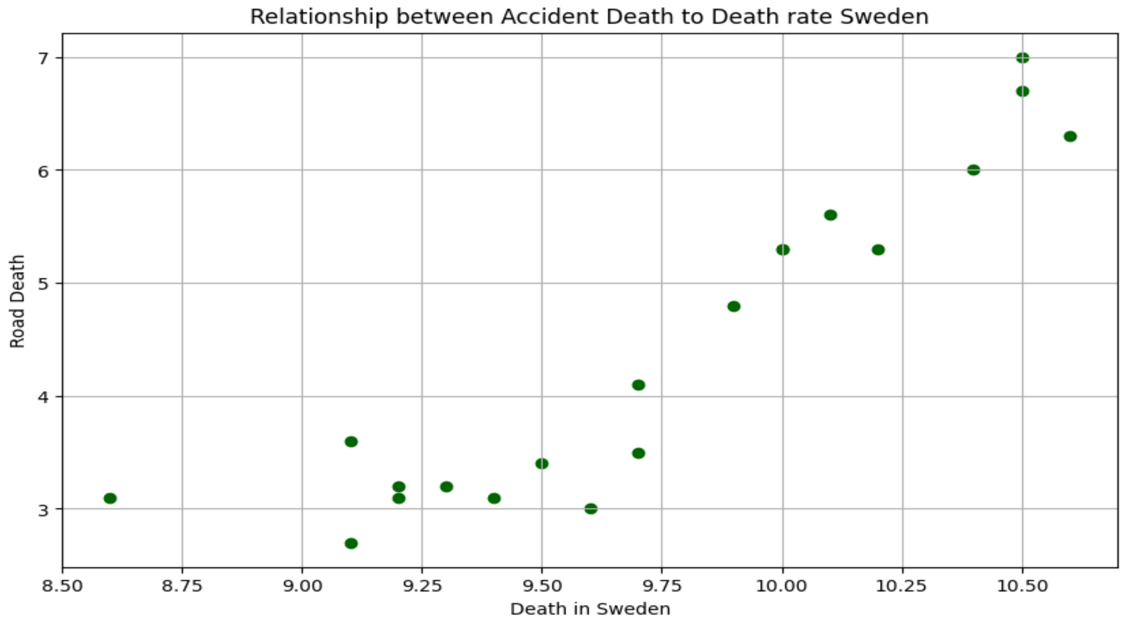


Figure 7 Scatter plot of Mortality Road accident vs Death rate in Sweden

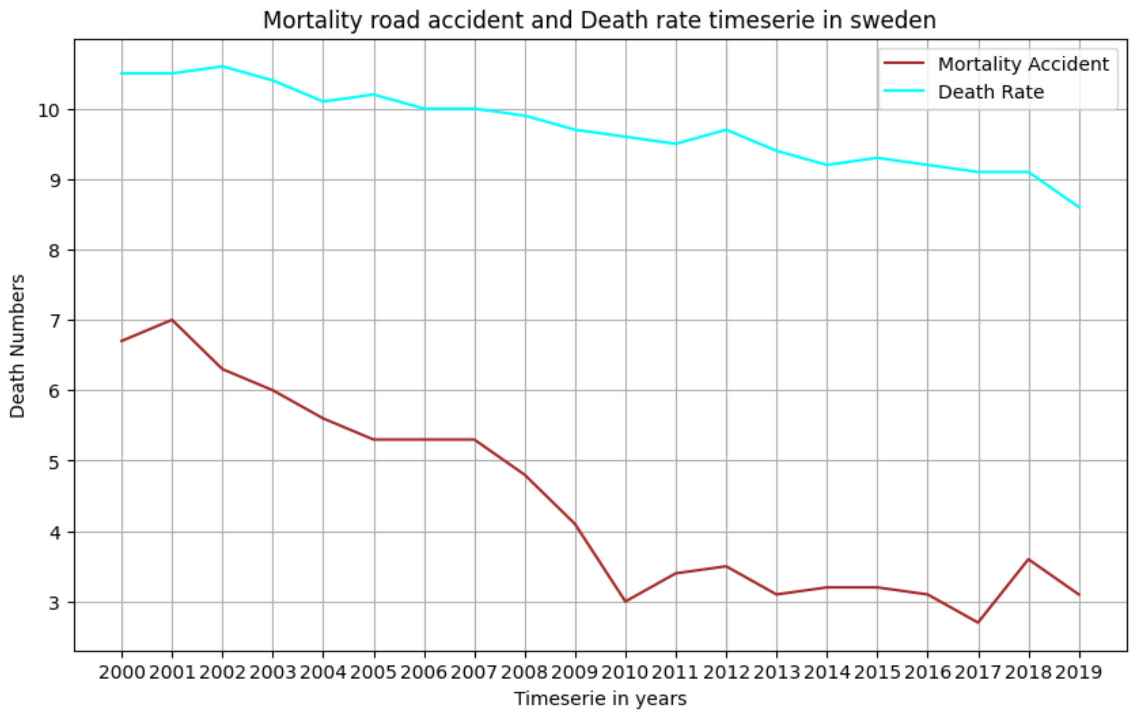
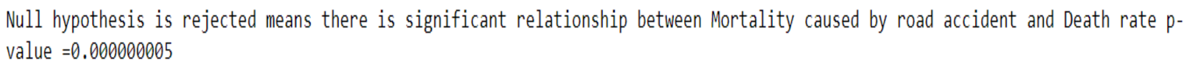


Figure 8 Mortality caused by accident and Death rate timeseries



```
Null hypothesis is rejected means there is significant relationship between Mortality caused by road accident and Death rate p-
value =0.000000005
```

The estimated Mortality caused by road accident in 2021 is 1.88. this shows that in Sweden in 2021 the estimated Mortality caused by road accident is 1.88 per 100,000 population. And the Average error for this prediction is 10.09% which indicate high level of accuracy in the prediction model. The Correlation coefficient is 0.91, this implies that the change in Mortality caused by road accident lead to the change in Death rate. For here in Sweden as the Mortality due Road accident decreases the Overall Death rate in Sweden decreases too. This is supported by graphs in the figure 7 the scatter plot shows that there is relationship between Mortality caused by road accident and Death rate in Sweden shows the trends, in the figure 8 shows the Mortality caused by road accident and Death rate against timeseries in the graph as the decrease in Mortality rate led to decrease in Death rate over years. Where in year 2000 the Mortality rate was very high and death rate too where. In year 2019 there is significant reduction in death caused by road accident and also decrease in death rate which is clear prove that that the prediction of Mortality caused by road accident reduced in year 2021 we predicted. As another evidence the hypothesis test say that p-value is 0.000000005 which is small p- value close to zero which help to reject the null hypothesis.

Conclusion

In Conclusion, the studying on mortality caused by road accident in Sweden shows the influence on overall nation Death rate. The analysis reveals that concerted efforts to reduce road accident-related mortality has resulted in significant decrease in the overall death rate. Furthermore, the study highlights how historical trends can be used in predicting the future trends and with low average error when I was predicting the mortality caused by road accident in year 2021. It is imperative for Sweden government in collaboration with its citizens, to prioritise road safety to improve their life health. By enforcing policy to reduce the road accident in country to ensure prosperous future for its population.

In this study I have improved the understanding the process of collecting data cleaning them and gain more insight in modelling and predicting method.

Q4.

In this question I was tasked to use published data by bank of Israel of unemployment rate (per 100 Israeli workforces) from specified period dec 1980 to Sep 2013 to estimate the rate of unemployment in 2020 and also evaluate the accuracy of the estimated value in percentage.

In addressing this question, I accessed the unemployment rate dataset for Israel through Quandl, a data platform. Accessing this data required the use of an API key, a unique identifier for each user, by using the imported quandl library I passed unique API key to retrieve the data. By using the quandl library, I streamlined the dataset by using the code of unemployment rate without need to download the data even if there is option is available.

By studying the extracted dataset, the period exceeds the requested period to use. I have filtered the code the dataframe to contain the data in period from 1980-12-31 to 2013-09-02.

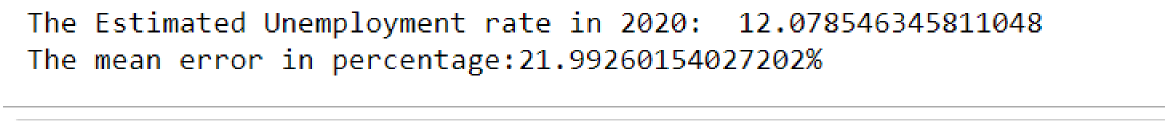
I have declared variable to store the predictor and predict data where predictor or independent variable is date and setting predict or dependent variable as the column of Unemployment values.

By using sklearn library I have declared variable to store instance of linear regression function, and also I have fitted the model by using fit() function to study the relationship between dependent and independent variable to find the best fitting linear equation based on the data given.

I employed a linear regression model to forecast the prospective unemployment rate in Israel for the year 2020. This projection was derived from examining historical data using regression models, and it was executed using the sklearn function predict() which help in estimation of the future scenario.

I have used Mean Absolute Percentage Error (MAPE) to determine how likely the predicted unemployment rate in 2020 could be accurate. In the code I have compare the actual value to estimated value and the value obtained determine the average error the estimation has.

I have printed the function that will return the estimated unemployment rate in 2020 and also average percentage error of this prediction.



```
The Estimated Unemployment rate in 2020:  12.078546345811048
The mean error in percentage:21.99260154027202%
```

As shown Above the estimated Unemployment rate in year 2020 to be 12.078 and other decimal and the error to this estimation by using MAPE is 21.993%. this shows the error which is not high based on historical data we are using stopped in 2013 which is 7 years to the estimated rate, mean if we use recent data can reduce the error significantly.

In conclusion by using linear regression to analyse the historical data gives a chance to estimate the change in future trends and then take action in this study I have learned that the more studying on historical data the recent one provide high degree of accuracy to its estimation.

Reference

[1] “scikit-learn: machine learning in Python — scikit-learn 1.3.1 documentation.” Accessed: Oct. 20, 2023. [Online]. Available: https://scikit-learn.org/stable/

[2] “stepwise-regression/stepwise_regression/step_reg.py at master · AakkashVijayakumar/stepwise-regression,” GitHub. Accessed: Oct. 22, 2023. [Online]. Available: https://github.com/AakkashVijayakumar/stepwise-regression/blob/master/stepwise_regression/step_reg.py

[3] Zach, “How to Calculate BIC in Python,” Statology. Accessed: Oct. 22, 2023. [Online]. Available: https://www.statology.org/bic-in-python/

[4] “What is R Squared? R2 Value Meaning and Definition,” freeCodeCamp.org. Accessed: Oct. 22, 2023. [Online]. Available: https://www.freecodecamp.org/news/what-is-r-squared-r2-value-meaning-and-definition/

[5] “World Bank Open Data,” World Bank Open Data. Accessed: Oct. 19, 2023. [Online]. Available: https://data.worldbank.org

[6] “World Bank Open Data,” World Bank Open Data. Accessed: Oct. 21, 2023. [Online]. Available: https://data.worldbank.org