Carnegie Mellon University Africa

# 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:**

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.

2) Steps:

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

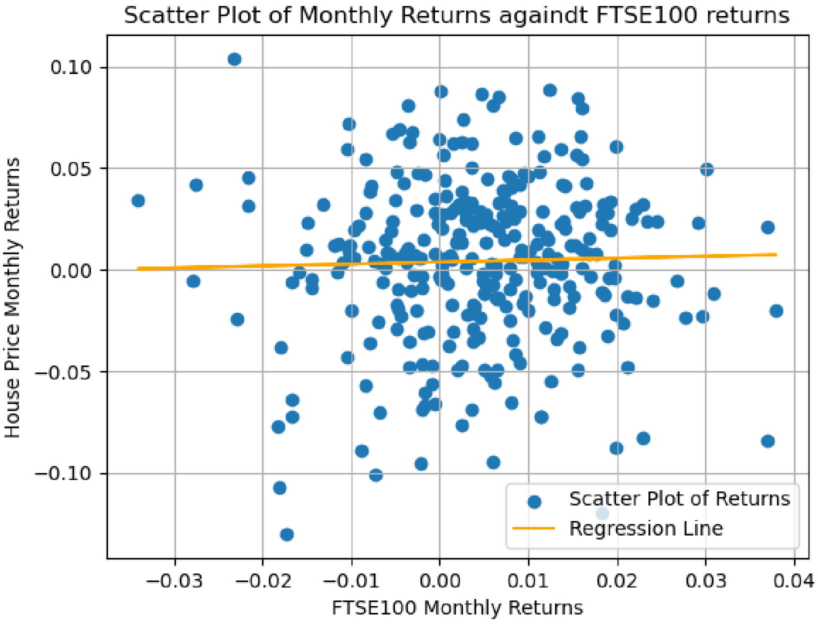
- The required data was imported.
- Fields, like the "Date", was renamed and converted to the appropriate date format.
- Monthly return for both data sets are computed using the *pct_change()* function.
- A regression model is created using *linregress* of the imported *stats* module.
- Using the intercept 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.

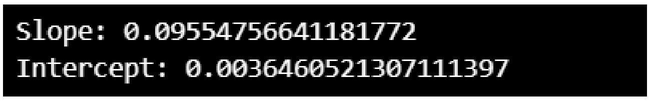
3) Results and observation:

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

a. 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





```
Slope: 0.09554756641181772
Intercept: 0.0036460521307111397
```

b. 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

c. Hypothesis test:

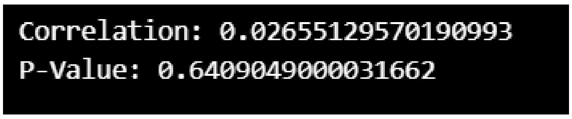
The null hypothesis, µ0:

- 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



```
Correlation: 0.02655129570190993
P-Value: 0.6409049000031662
```

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.

2) Steps:

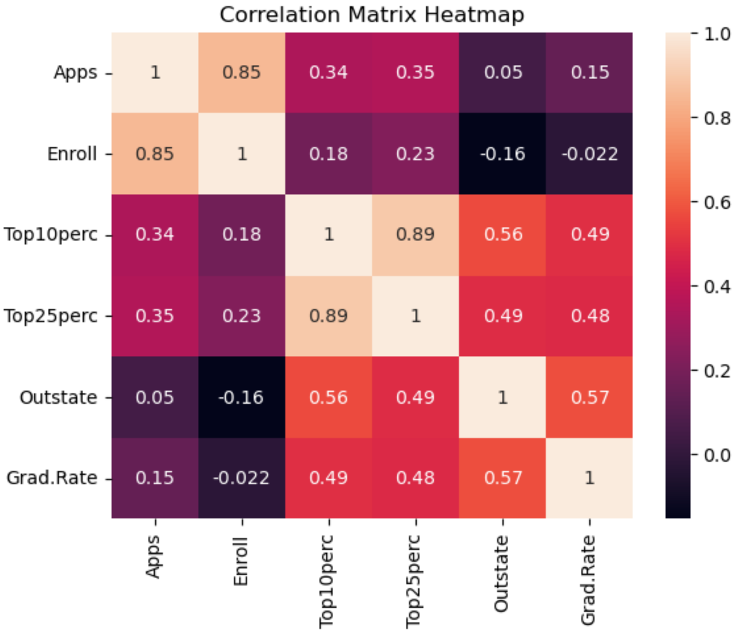
To arrive at the results and insights:

- 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.

3) Results and observation:

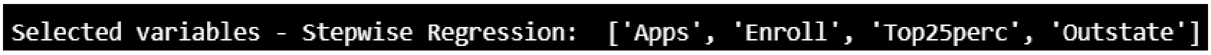
Bellow are the results and observations achieved.

a. The correlation between the variables can be represented using the heat map bellow:



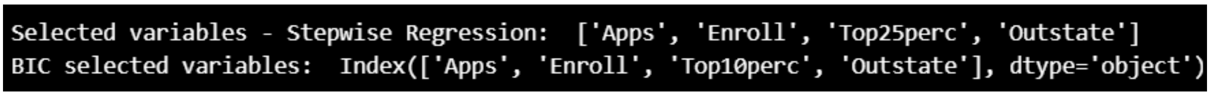
b. A stepwise backward regression was performed to select predictors. In this process, we commence will all independent variables, and by fitting the model to each feature, we eliminate the least significant values using the p-value threshold of 0.05, till all features are statistically significant. To ease this process, we used a function proposed by Askkash[4].

c. Using backward regression, we fall on the following predictor variables: Apps, Enroll, Top25perc, and Outstate



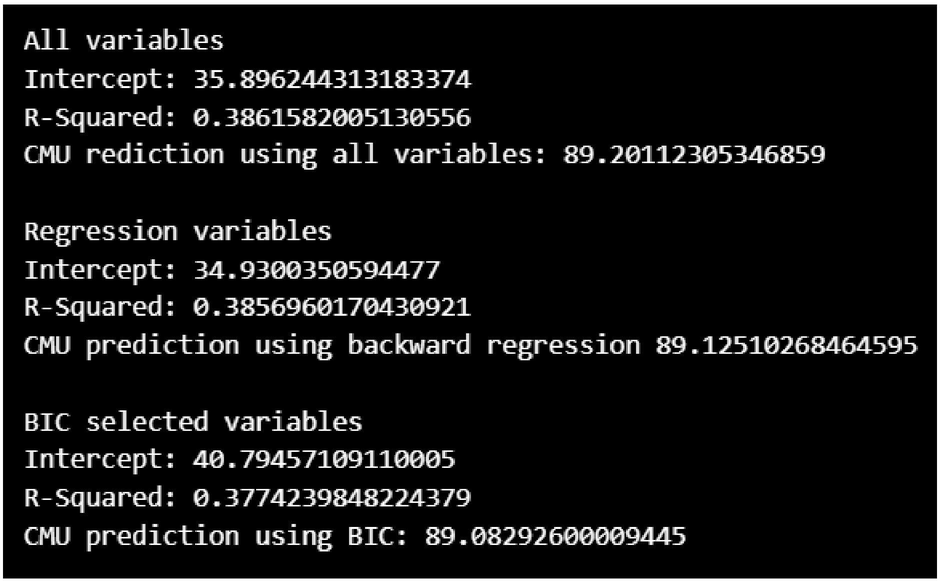
```
Selected variables - Stepwise Regression:  ['Apps', 'Enroll', 'Top25perc', 'Outstate']
```

d. 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.



```
Selected variables - Stepwise Regression:  ['Apps', 'Enroll', 'Top25perc', 'Outstate']
BIC selected variables:  Index(['Apps', 'Enroll', 'Top10perc', 'Outstate'], dtype='object')
```

e. 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.



```
All variables
Intercept: 35.896244313183374
R-Squared: 0.3861582005130556
CMU rediction using all variables: 89.20112305346859

Regression variables
Intercept: 34.9300350594477
R-Squared: 0.3856960170430921
CMU prediction using backward regression 89.12510268464595

BIC selected variables
Intercept: 40.79457109110005
R-Squared: 0.3774239848224379
CMU prediction using BIC: 89.08292600009445
```

f. 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.

2) Steps:

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

a. Mortality caused by road traffic injury (per 100,000 population)
b. 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.

3) Results and observation:

We arrived at the following:

a. Hypothesis testing

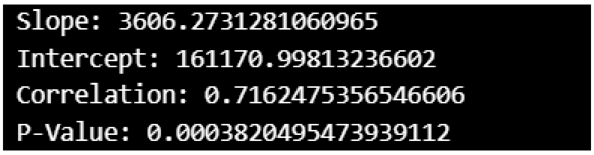
The null hypothesis, µ0:

- 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.

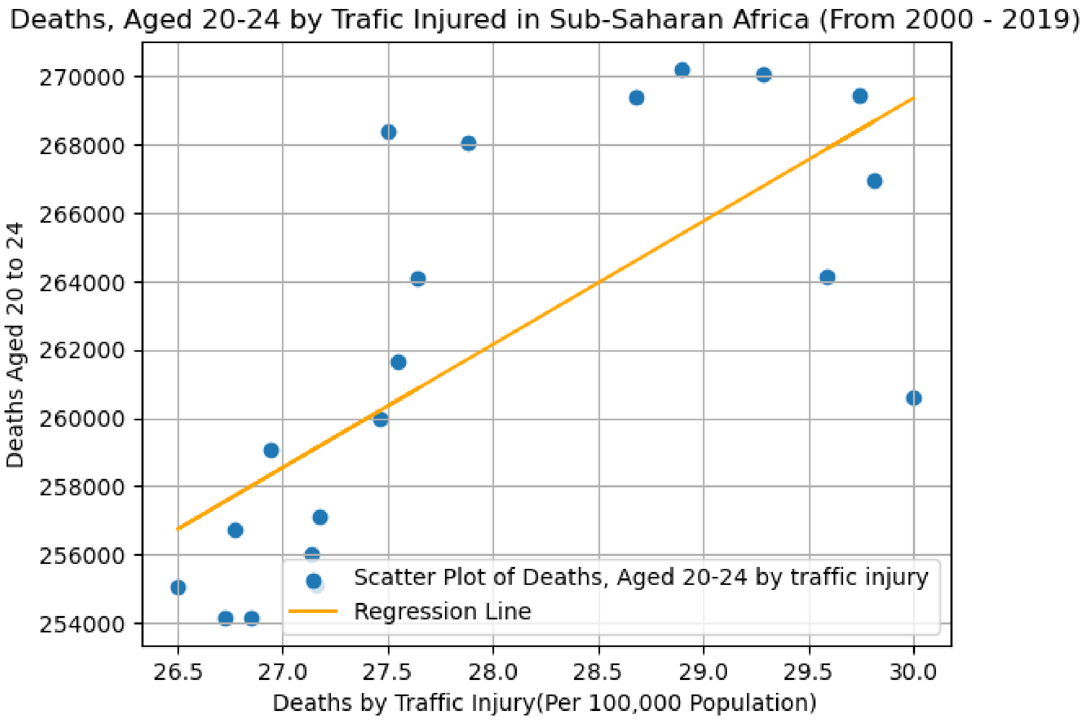
With α = 0.05 we perform a Pearson test of correlation the following p-value was observed. With a P value of 0.00038 << α = 0.5, we conclude that there is a significant relationship between both values. For emphasis, a one tailed test is performed since we want to check for a positive correlation only.



```
Slope: 3606.2731281060965
Intercept: 161170.99813236602
Correlation: 0.7162475356546606
P-Value: 0.0003820495473939112
```

b. Plots:

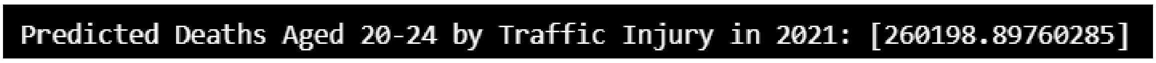
The following scatterplot of both variable and line plot of the model was obtained. We observe a strong correlation between both variables as increased in death by traffic injury per 100,000 is reflected in an increase in deaths of 20 – 24-year-olds, with a slope of 3606, and an intercept of 161170.



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.

c. 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.



```
Predicted Deaths Aged 20-24 by Traffic Injury in 2021: [260198.89760285]
```

**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.

2) Steps:

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

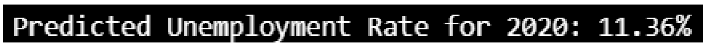
- Dataset is downloaded and extracted into data frame.
- Data from 1980 to 2013 is extracted and the date is converted to ordinal using the toordinal() python function.
- A linear regression model is fitted using date as independent variable and unemployment as explanatory variable.
- We predict the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

3) Results and observation:

We observed the following.

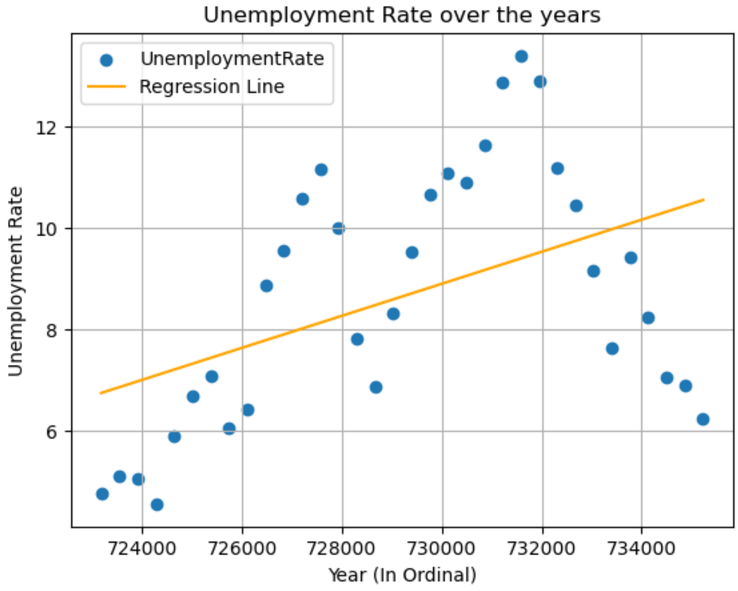
a. Model and prediction

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.



```
Predicted Unemployment Rate for 2020: 11.36%
```

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

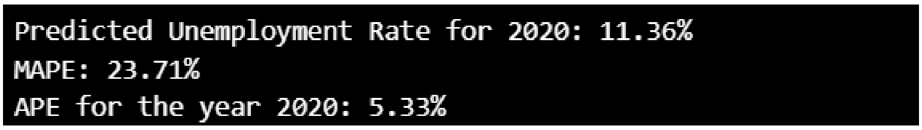


b. Accuracy evaluation:

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.

$$MAPE = \frac{1}{n}\sum_{1}^{n}\left|\frac{Ai - Fi}{Ai}\right| \times 100$$

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.



```
Predicted Unemployment Rate for 2020: 11.36%
MAPE: 23.71%
APE for the year 2020: 5.33%
```

## 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