CARNEGIE MELLON UNIVERSITY - AFRICA

DATA, INFERENCE & APPLIED MACHINE LEARNING

(COURSE 18-785)

Professor Patrick McSharry

Assignment 4

MUNYANEZA Kenny Roger

23nd October 2023

Introduction

This is a final report designed as a fulfillment of the requirements to complete assignment 4 DIAML.

Tools used

- Python
- Vs Code

Question 1

Third party libraries used

- **pandas**: It was used to read the Stock exchange file and the House price dataset and manipulate them
- **scipy**: Used the stats submodule to build a linear regression model given the dependent and independent variables.
- **matplotlib:** Used the Pyplot submodule to plot the graphs
- **Statsmodels:** Used to calculate a linear least-squares regression for two sets of measurements.

Implementation process used

To solve this problem, I read the two datasets into separate dataframe and calculated monthly returns for each dataset. I assigned monthly returns of each dataset in two objects, and I passed those objects as arguments to the **"linregress()"** function to get all the attributes that describe the linear regression model with Stock Exchange as dependent variable and House Prices as independent variable. The attributes are the slope, the constant, the correlation coefficient, the p-value, the standard error, and the standard error of the estimated intercept. Next, I fitted the model using the **Ordinary Least Squares (OLS)** method of linear regression, and I made prediction of values using the "**.predict()"** function which gave me values predicted by the model.
Lasty, I plotted the given data values of Stock Exchange against House Prices, I drew the line best fit of the model, and I scatter plotted the predicted. I plotted all this information using a matplotlib scatterplot.

Results

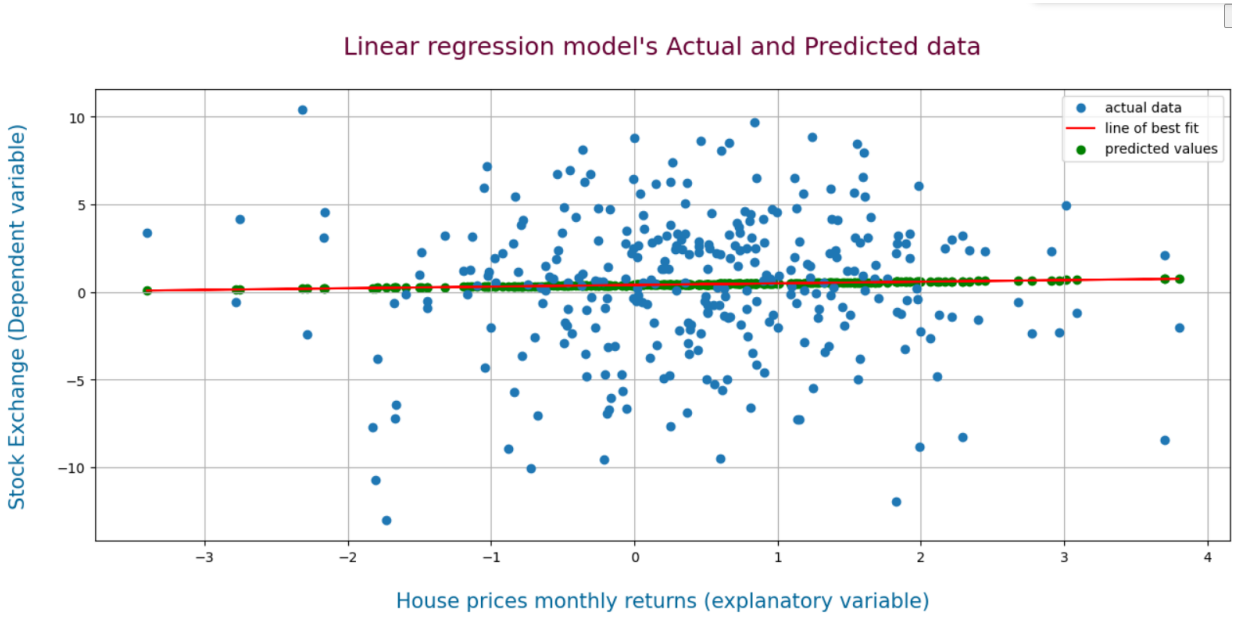
```
Slope (Beta): 0.09324142754349976
Intercept (Constant): 0.4047837686662456
Correlation Coefficient (R): 0.02655129570190993
P-Value: 0.640904900003165
```

Interpretation of the results

- Since the slope is positive, the Stock Exchange and the House Prices are directly proportional.
- Since the constant is 0.40, when the House Price is zero the Stock Exchange values is equal to 0.40.

- The value of the correlation coefficient indicates that there is a weak negligible linear relationship between House Prices and Stock Exchange, which means that House Prices are not a powerful or relevant predictor of Stock Exchange (FTSE100).

Graph of the Predicted and actual values of monthly returns



Interpretation

The graph is highly scattered which affirms the weak relationship between House Prices and Stock Exchange (FSTE).

Hypothesis Testing

- **Null Hypothesis:** The slope is zero and there is no significant relationship between Stock Exchange and House Prices.
- **Alternative Hypothesis:** The slope is different from zero and there is a significant relationship between Stock Exchange and House Prices.

Result

```
P-Value: 0.640904900003165
Failure to reject the null hypothesis: There is no significant relationship between house returns and FTSE100 returns.
```

Since the p-value is greater than the significant level 0.05, we fail to reject the null hypothesis, hence there is no significant relationship between Stock Exchange and House Prices as it was concluded in earlier interpretation of the results.

Question 2

Third party libraries used

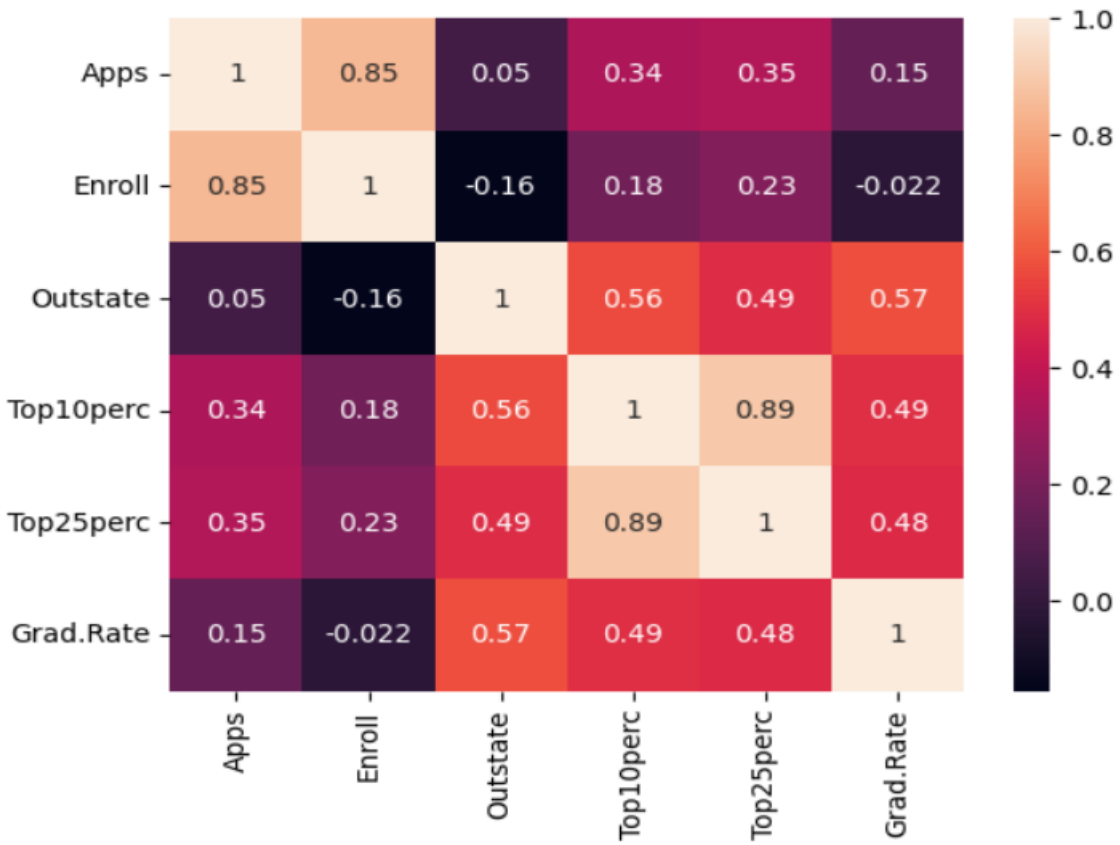
- **pandas**: It was used to read the Stock exchange file and the House price dataset and manipulate them
- **numpy:** Used create and manipulate arrays in this problem.
- **Stepwise_regression:** Used to perform backward / forward regression.
- **Seaborn:** Used to create informative statistical graphics.
- **Sklearn:** Used train, build, test and evaluate machine learning models.

Implementation Process

To solve this problem, read the College.csv dataset using pandas, extracted the necessary columns of the independent variables, namely "Apps", "Enroll", "Outstate", "Top10perc", "Top25perc" and the dependent variable "Grad.Rate".

Then, I used the corr() function to get a matrix of correlation coefficients amongst predictor variables and between predictor variables and the target variable.

Correlation coefficient matrix



- The correlation coefficient between Apps and Grad.Rate is 0.15, which indicates a weak positive linear relationship between the two variables.
- The correlation coefficient between Enroll and Grad.Rate is -0.022, which indicates a weak negative linear relationship between the two variables.
- The correlation coefficient between Outstate and Grad.Rate is 0.57, which indicates a strong positive linear relationship between the two variables.
- The correlation coefficient between Top10perc and Grad.Rate is 0.49, which indicates a moderate positive linear relationship between the two variables.
- The correlation coefficient between Top25perc and Grad.Rate is 0.48, which indicates a moderate positive linear relationship between the two variables.

Next, I built a linear regression model using stepwise, considering the Grad.Rate column as the dependent variable. I performed **forward_regression** on the independent variable of the model to get the most useful predictor variables, I used BIC selected model to see the predictor variables that would be useful in that case, I compared the accuracy of the three models namely the model with all predictor variables, the BIC model, and the model obtained by forward_regression.

Results

- ### Summary statistics of the linear regression model with all predictor variables

OLS Regression Results

Dep. Variable:	Grad.Rate	R-squared:	0.386
Model:	OLS	Adj. R-squared:	0.382
Method:	Least Squares	F-statistic:	97.00
Date:	Mon, 23 Oct 2023	Prob (F-statistic):	2.73e-79
Time:	07:01:20	Log-Likelihood:	-3121.9
No. Observations:	777	AIC:	6256.
Df Residuals:	771	BIC:	6284.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	35.8962	2.137	16.800	0.000	31.702	40.091
Apps	0.0008	0.000	3.005	0.003	0.000	0.001
Enroll	-0.0030	0.001	-2.828	0.005	-0.005	-0.001
Outstate	0.0017	0.000	11.124	0.000	0.001	0.002
Top10perc	0.0493	0.065	0.762	0.446	-0.078	0.176
Top25perc	0.1813	0.055	3.312	0.001	0.074	0.289

Omnibus:	23.498	Durbin-Watson:	1.946
Prob(Omnibus):	0.000	Jarque-Bera (JB):	41.256
Skew:	0.210	Prob(JB):	1.10e-09
Kurtosis:	4.048	Cond. No.	5.13e+04

- Useful variables in predicting using stepwise regression

The useful variables are **Outstate** and **Top25perc**. I obtained these results by calling the "forward_regression" with an object of predictor variables as first parameter, an object of dependent variable as second parameter, the significance level (0.05), and a Boolean value for the last parameter called verbose that specifies the logging of the process. The "forward_regression" is obtained from the external library called **stepwise_regression.**

- Would the above variable still be useful if BIC were used to select the model?

Yes, the same variables **Outstate** and **Top25perc** would still be useful if BIC were used to select the model because from my computation to determine the most parsimonious model with the lowest BIC value possible is made of a combination of **Outstate** and **Top25perc.**

- Compare the accuracy of the model using only useful predictors and the model using all five predictor variables.

```
R-squared (BIC Model): 0.4834179169125846
R-squared (Stepwise Model): 0.4834179169125846
R-squared (For all predictors): 0.4528562161254003
Mean Squared Error (BIC): 121.50146022950811
Mean Squared Error (Stepwise): 121.50146022950811
Mean Squared Error (For all predictors): 128.68965237614097
```

The BIC model and Stepwise model have the same accuracy given that their Coefficient of Determination (r^2) and Mean Squared Error (MSE) are equal. Also, the model using all five predictor variables is less accurate than the BIC model and the Stepwise model given that its MSE is higher than that of the two other models and the r^2 of the model with 5 predictor variables is lower than that of the two other models. [1]

- Prediction of the Graduation Rate of Carnegie Mellon University

```
The graduation rate at CMU would be: [86.83968987]
```

Conclusion:

The predicted Graduation differs significantly from the actual Graduation Rate of CMU, which is 74. This is mainly because all predictor variables have a weak correlation with the dependent variable (Graduation rate). Therefore, to get a prediction that is like the actual graduation rate, better predictor variables with a stronger correlation with the graduation rate would be used.

Question 3

Third party libraries used

- **pandas**: It was used to manipulate the data-frame provided which involved reading the dataset in csv format and calculating the correlation coefficient.
- **Statsmodels:** Used to calculate a linear regression model.

Problem Statement

My study is about the trend in the increase of Fossil fuel energy consumption in **Turkey** with relationship to the growth in export of goods through railway transport in that country.

Source of data

I obtained two datasets from the world bank indicators.

- The first dataset is a distribution of exportation by railway measured over the years in million ton-Kilometer. This unit of measurement is commonly used in the context of transportation and logistics. A "million ton-kilometers" means that one million metric tons of cargo have been transported over one kilometer or its equivalent in kilometers.
  The dataset can be accessed here.
- The second dataset is a distribution of Fossil fuel energy consumption as a percentage of the available energy. It can be accessed here.

Assumptions

- Fossil fuel energy consumption as a percentage of the available energy has increased over the years.
- The increase in exportation by railway transport has directly impacted the increase in fossil fuel energy consumption.
- The future consumption of fossil fuels can be predicted using expected measure of exportation by railway transport.

Methodology used

I extracted the measurements for Turkey from both the dataset for Fossil fuel consumption and the dataset for exports through railway transport over the years 1995 to 2015. Transposing each row gave me a two column dataframe of year and value columns. **I used exports through railway transport, and years as independent variables** and **I used fuel consumption as dependent variable**.

I created a linear regression model that uses rail transport exportation, and years as predictor variables and fuel consumption as target variable using Ordinary Least Squares (OLS) method of linear regression.

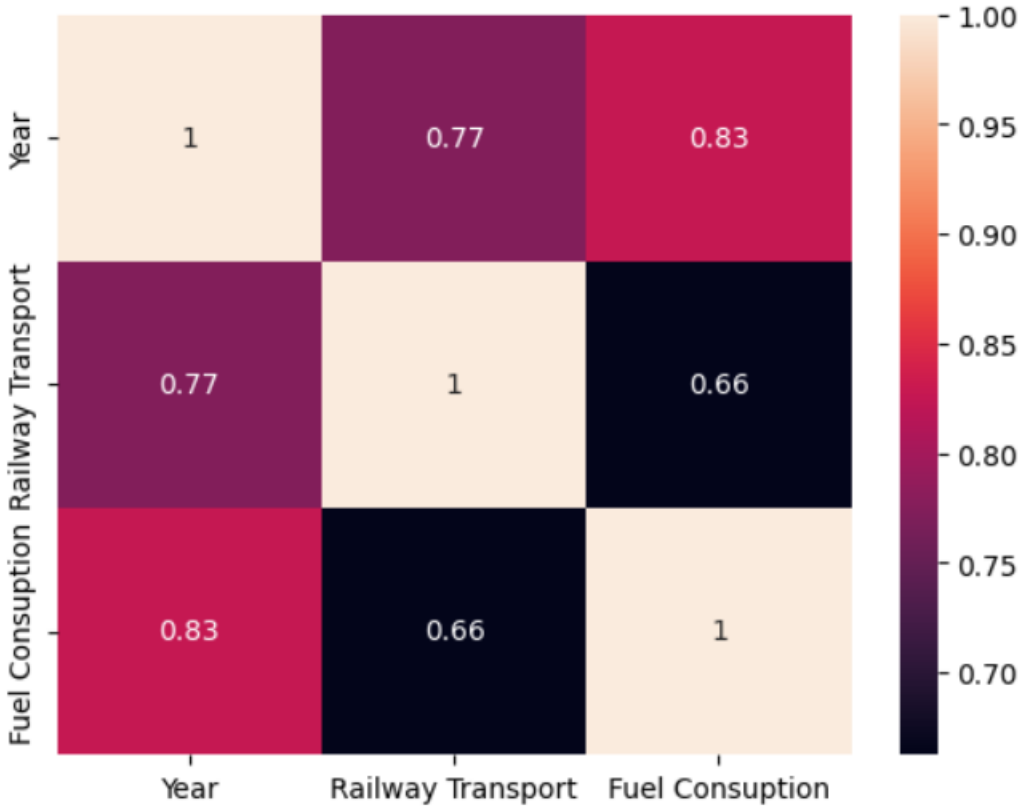
Statistical Analysis

- The Mean Squared Error is 2.26.
- The Coefficient of determination is 0.68

```
Fuel Mean Squared Error (MSE): 2.2585265343028746
Fuel R-squared (R²): 0.6834869629694473
```

These statistics suggest that the model built has moderate accuracy of about 68% and moderate level of error which is 2.26.

- The coefficient matrix of predictor variables and the dependent variable:



From the figure above, there is a strong linear correlation of 0.83 between Fuel Consumption and Years. Also, there is a moderate strong linear correlation of 0.66 between the Fuel Consumption and Railway transport.

Prediction

To predict the situation in 2021, I used the model predict the fuel consumption in case the year was 2021 and the railway transport exportation was 16300 million ton-kilometers.

The predicted value of fossil fuel consumption in 2021 is: 93.00 % of the available fossil fuel energy.

```
The Fuel consumption in 2021 would be: 93.0031439803763
```

Conclusion

- Given the statistics of the model, the years passed, and rail transport exportation can be used to predict fossil fuel consumption.
- The predicted value is consistent with the assumptions made because it is higher than the value of fossil fuels consumption in past years.

Question 4

Third party libraries used

- **Quandl**: It was as a source of the dataset used in this question.
- **pandas**: It was used to manipulate the data-frame provided which involved reading the dataset in csv format.
- **Statsmodels:** Used to calculate a linear regression model.
- **Sklearn:** Used train, build, test and evaluate machine learning models.

Implementation Process

Using the dataset provided I created a linear regression model that predicts unemployment rates using the year as the predictor variable. I first filtered out the dates before or beyond the required timeframe (1980-12-31 to 2013-09-02). Then, I appended a column of years corresponding to each date to the dataframe which I used as independent variable. Lastly, I fitted the linear regression model predicted the unemployment rate for the year 2020.

Results

```
Unemployment prediction in 2020: 12.078546345811276
```

Accuracy of the model

To determine the accuracy of the model I used the fitted linear regression model to predict all the unemployment rates from 1980 to 2013 inclusive. I then used the formula of Mean Absolute Percentage Error (MAPE) using an array of actual unemployment rates and an array of predicted unemployment rates using the formula below. [2]

$$\textbf{MAPE} = (1/n) * \Sigma(|\text{actual} - \text{prediction}| / |\text{actual}|) * 100$$

MAPE value Obtained

```
The accuracy of the mode determined by MAPE is 21.99260154027202
```

The value obtained implies that the predictions of the model created deviate from the actual values by 21.99 %. This means that the model has a low accuracy of predicting the unemployment rate.

[1] ‘IQCode - Learn to code’, IQCode.com. Accessed: Oct. 23, 2023. [Online]. Available: https://iqcode.com/code/python/IQCode.com
[2] Zach, ‘How to Calculate MAPE in Python’, Statology. Accessed: Oct. 23, 2023. [Online]. Available: https://www.statology.org/mape-python/