DOI: 10.46541/978-86-7233-406-7_228

27th International Scientific Conference
Strategic Management
and Decision Support Systems
in Strategic Management
SM2022
Subotica (Serbia), 20th May, 2022

Dijana Jovanoska
PhD student
University St. Kliment Ohridski Bitola,
North Macedonia
dijana_67bmis@yahoo.com

Gjorgji Mancheski
Full proffesor
University St. Kliment Ohridski Bitola,
North Macedonia
gmanceski@t-home.mk

ON-LINE BIG DATA PROCESSING USING PYTHON LIBRARIES FOR MULTIPLE LINEAR REGRESSION IN COMPLEX ENVIRONMENT

Abstract: The phenomenon called Big Data today is one of the most significant and least visible consequences of the development of technology and the Internet. Namely, the data generated by today's globally connected world is growing at an exponential rate and they are a real "gold mine" for those users who know how to correctly interpret such data and make successful decisions based on them. Data analysis and processing is one of the most important components of a large data system, and in this branch of data science the most popular is the Python programming language, which provides its users with a large number of constantly maintained program libraries and developing environments. The most important thing for legal entities and individuals is that almost all program libraries and functions provided by this programming language come with free licenses and possess open code, maintained and quality technical documentation, which provides each company with significant money savings and time.
This research paper is dedicated to the possibility of determining and creating a multi regression model of large amounts of data by using Python, on the basis of large amounts of data provided by two market retailers in order to display a multi regression model and assess its predictive power. Because the number of variables is large, several models have been made in this research paper and a comparative analysis of the different models has been made, which shows that Python is a good tool that can be used repeatedly to select different variants and evaluate the resulting model for which a graphical interface can be made and would be much more acceptable as an end user, can be placed on a server on the Internet or on a modern Cloud platform and used by users as an on-demand concept and the results can be embedded in end-user interfaces and models made in this way (with dynamic data extraction)can be used in BI and machine learning processes.
Keywords: Python, big data, data processing, multiple linear regression.

1. INTRODUCTION

The subject of this research paper is a multiple regression model based on a large amount of data provided by two retail markets. The data available from these two retail markets comes with a large number of rates from the retail itself. The data taken for processing are from 2019. The independent variable in the model is the difference in prices achieved by the company from sale, and the dependent variables is the purchase /the sale value of the goods that are procured from the suppliers. The number of suppliers to the retail markets is 141. Since the number of suppliers is very large, it is obvious that the multiple regression model would contain many variables and a way should be found to reduce it.
The research methodology used in this paper is to create a multiple regression model and evaluate its predictive power. Because the number of variables is large, several models have been made in the paper and a comparative analysis of the different models has been done.

The purpose of this research is to form a regression model that will determine the differences in prices which are in function of the amount of the purchas/the sale value of goods from suppliers. The possibility of using it to predict future price differences will also be checked. The independent and dependent variables vary within a week over a year.

2. MULTIPLE REGRESSION MODEL

The linear regression model is very often applied in processes where linear dependence can be detected. The number of independent variables certainly depends on the problem being researched and the goal we want to achieve. The simplest regression model is[1]:

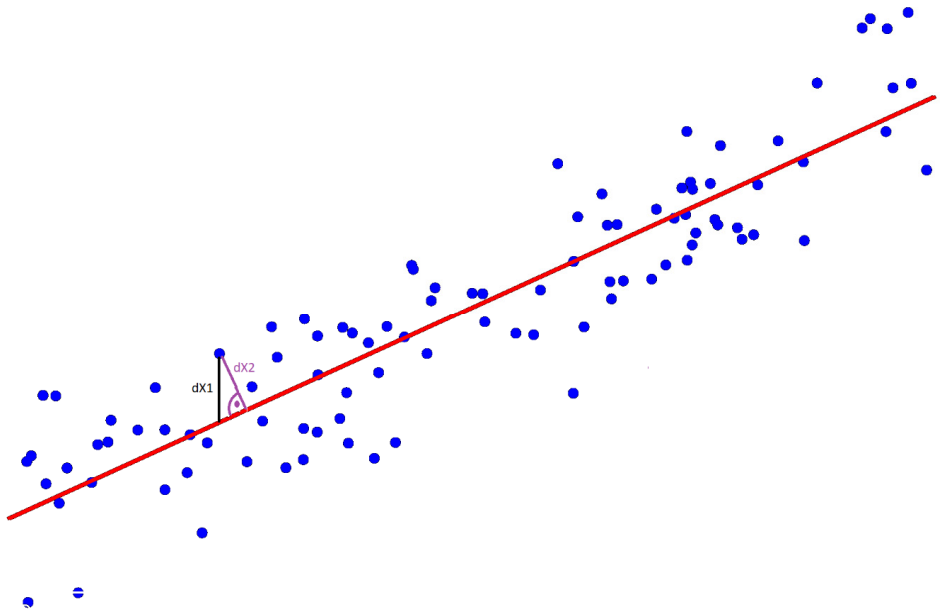
$$y = \alpha + \beta x$$

The dependent variable y is in function of the independent variable x. There is no process that can cover all the independent variables that affect the dependent variable. However, to make the model with some estimated error, we introduce a variable ε so the model looks like this:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

This is the model for one-dimensional regression model.

His graphic representation would be:



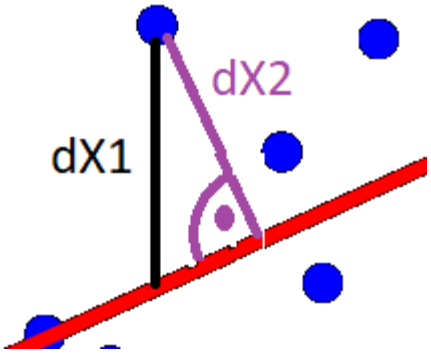
Picture 1: Graphic representation of a simple linear model
Source : https://en.wikipedia.org/wiki/Simple_linear_regression#/media/File:Linear_regression.svg (modified)

There are two ways to determine the linear regression **line**

1. Optimization (minimization) of vertical differences from individual results to imaginary regression line (stable multiple linear regression). This is done by minimizing the sum of the squares of these distances dX1 to Picture 1.

2. Optimization (minimization) of the normal distances from the individual results to the imagined regression line (stable multiple linear regression). This is done by minimizing the sum of the squares of these distances dX2 to Picture 2.

[1] Taha, Hamdy A.,Operations Research: An Introduction 8th ed., Pearson Prentice Hall, 2007



Picture 2: Excerpt from Picture 1

But the processes in reality almost never depend on only one variable and that is why the most practically applied model is the multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

In order to be able to make such a model, it is necessary to have an appropriate set of data that we can mathematically represent with[2]:

$$\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$$

The vector record of the regression model is:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

Where the vector Y is the vector of the dependent variables, the vector X is the vector of the independent variables.

The vector $\boldsymbol{\beta}$ is the vector of the coefficients before the independent variables, and in the end $\boldsymbol{\varepsilon}$ is the vector of the estimated model errors.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

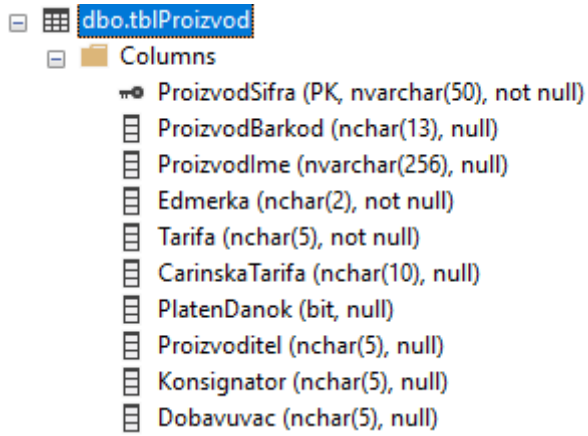
Where the vector Y is the vector of the dependent variables, the vector X is the vector of the independent variables.

The vector $\boldsymbol{\beta}$ is the vector of the coefficients before the independent variables, and in the end $\boldsymbol{\varepsilon}$ is the vector of the estimated model errors.

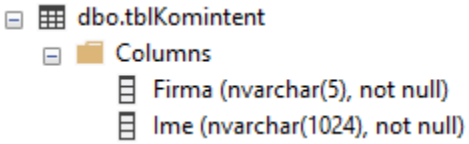
3. DESCRIPTION OF THE DATA AVAILABLE

[2] Damadar. N.Gujarati, *Basic Econometrics*, Fourth Edition, Tata McGraw Hill, 2004

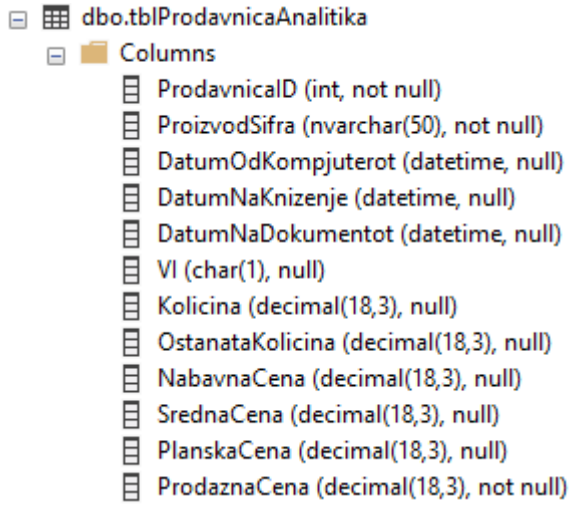
The available data belongs to a company that has two retail markets. They were available to me on a server set up on the Internet that has an MSSQL server installed. The structure of the products spreadsheet (tblProizvod) had 96 fields. The analysis showed that only two of them are necessary for this research, and they are (ProizvodSifra, Dobavuvac)



From the table of clients **(tblKomintent)** which also has about 100 fields, only two are necessary:



The table **(tblProdavnicaAnalitika)** also has about 100 fields, and only some were necessary, as follows: (ProizvodSifra, DatumOdKompjuterot, VI, Kolicina, NabavnaCena, SrednaCena, ProdaznaCena), is:



The data were unprocessed and uncategorized. The database contains 2336 products. The total number of suppliers is 141, which is a very large number as independent variables, so in the modeling process some suppliers were rejected for which there is a insignificant amount of procurement throughout the year. First, the elimination was made for companies that have smaller purchases than 60,000 denars. With such a move the number of variables was reduced to 119. Consciously with such a procedure it increases the error in predicting the model. The next step was to eliminate the companies for which the total procurement is less than 200,000 denars. With such a move, the number of independent variables was reduced to 92. Since the number was large again, in the end I reduced the number of variables to companies whose total amount of procurement is 500,000 denars. If we take into account that the total purchases of the company are 244,648,012 million denars, the amount of 500,000 denars is only 0.2% and in my opinion it is justified. Of course, this refers to one company, and in the model, however, several companies are eliminated. Doing such an elimination, procurements in the amount of 10,785,225 denars were eliminated. In percentage terms, it is 4.41 %, which is certainly acceptable as a built-in statistical error of the model.

As a result, a restriction has been accepted that the model will include suppliers who procured more than 500,000 denars during the year.

Of course, the software that was made has this data as a parameter and they can be easily changed and the model can be recalculated.

The second problem was the extraction of data from a large database where using a public network (Internet).

The code is divided into 5 parts depending on the purpose for which it is written. Part of the code that was made for database analysis and data extraction in Python can be found at the following link https://docs.google.com/document/d/1GLvm__OjpjjAauIvi5Xns6WGHmEWKz0t/edit?usp=sharing&ouid=117631998925527807241&rtpof=true&sd=true

The results of this code are stored in several files which represent input data for the next code that determines the regression model.

Those files are:

Total turnover

VkupenPromet.txt

(Decimal ('244648012.606599'), Decimal ('294747248.433000'), Decimal ('50099235.826401'))

Procurement of suppliers by weeks

RezultatiPoNedeli.csv

Value of independent variables (csv file – comma separated):

F00001, F00028, F00012, F00031, F00043, F00066, F00010, F00070, F00049, F00118, F00249, F00037, F00243,

...

20901,39139,20971,8940,9601,13162,16592,16912,11867,7851,8118,0,23377,8002,12573,4065,152,15574,9386,16389,20233,7380,6328,5317,5340,10873,7279,0,8163

Cumulative results by weeks.

RezultatiPoNedeli.csv

Value of independent variables (csv file – comma separated):

Nabavki,Prodazbi,RazlikaVoCeni
3702023,4457730,755706

Companies used as independent variables

CompaniesForPrediction.txt

Suppliers from which a significant procurment[3] was made during the year (JSON object file):


```
[
  {
    "Amount": 0,
    "CompanyID": "00001",
    "CompanyName": "ВИКТОРИЈА ТОБАКО ДООЕЛ",
    "TotalAmount": 1065914
  }
  {
    "Amount": 0,
    "CompanyID": "00028",
    "CompanyName": "СТЕФАЛЕК ЗД",
    "TotalAmount": 272662
  },
```


[3] For companies with significant procurements are taken into account those where the company procured more than 500,000 denars during the year

...


```
  {
    "Amount": 0,
    "CompanyID": "00151",
    "CompanyName": "ПАНОРАМА ДООЕЛ ОХРИД",
    "TotalAmount": 9823
  }
]
```


The data in this (JSON file) is as follows:

CompanyID – a symbol found in the independent variables with F as a sign. CompanyName – unnecessary for regression analysis but important for data comprehensibility.

TotalAmount – average value of weekly purchases throughout the whole year.

Amount – An amount that we can change when predicting by the model, and this is the average amount of procurement from the supplier on a weekly basis.

We did the testing and predicting of the model by changing this amount as an input parameter in the model.

4. RESULTS OF THE SOFTWARE SOLUTION AND THE MODEL

In order to determine the multiple regression model We wrote the following code in Python:

```
import json
from codecs import open

import pandas
from sklearn import linear_model

dx = pandas.read_csv("RezultatiPoNedeli.csv")
dy = pandas.read_csv("KumulativniRezultatiPoNedeli.csv")

X = dx
y = dy['RazlikaVoCeni']

regr = linear_model.LinearRegression()
regr.fit(X, y)
print(regr.coef_)

with open('CompaniesForPrediction.txt', 'r', encoding= 'utf-8') as companyForPrediction:
    data = json.load(companyForPrediction)

predictArray = []
for com in data:
        predictArray.append(com['Amount'])

print(predictArray)

RazlikaVoCeniPredviduvanje = regr.predict([predictArray])

print(RazlikaVoCeniPredviduvanje)
```

As a result of using this software solution We got the following regression coefficients:

```
[ 0.34232206  0.36901386  0.03947287 -0.4011094  -0.56947859  0.53512375
  1.26373386  0.26306351  0.49026068 -0.05833276 -0.49519478  0.65198297
  1.53424374  0.7022764  -0.65194589 -0.06106488 -0.66557829  0.20908608
 -0.25951186 -0.03081467  0.38428972  0.15261404  0.42047392  0.05031037
 -0.72902956 -0.075895    0.02832403  0.96316697  2.52303939  0.8985159
 -0.18227606 -0.41584122 -0.24450042  1.60245546 -0.62053319  0.21672189
```

0.06783619 -0.19052886 -0.01674749 0.51796349 -0.24170726 1.07363384
1.06600022 -1.97282974 0.74374664 0.21710738 -0.48849358 -0.65333522
0.27503429 0.65349738 0.56468211 0.59565353 0.98788953 2.43411822
-0.96790626 -0.28574629 0.2331749   1.57563868 0.72288802 0.79199459
-1.89653719 -0.22707687 -1.49194058 0.45711016 0.18773113 0.02548438
0.12700764 0.17251481 0.33364754 1.77102672 0.04983566 0.89410589
0.39294336 -0.11680222 0.55803886 0.74266238]

We made several predictions with the model and got the following results:

Predictions:

Prediction 1

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[-66,955.3145146]

In the first prediction We assumed that the firm was not making any purchases and as a result We got the axis section of the dependent variable α = -66,955.

Then We made two more predictions, Prediction 2 and Prediction 3:

Prediction 2

[1065914, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[29,7930.55920969]

Prediction 3

[0, 272662, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[33,660.74203737]

In order to be able to synthesize the interaction for both suppliers We made a third prediction as a synthesis of the previous two and the following was obtained:

Prediction 4

[1065914, 272662, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[39,8546.61576166]

This model can of course be used for prediction with more variables and for that you need to fill in the Amount fields in the file named *CompaniesForPrediction.txt*.

5. CONCLUSION

While researching the possibility of creating a multiple linear regression model by using Python We can summarize the following:

- We structured the selected model for linear regression analysis depending on the target function (the dependent variable we want to define);
- We made an analysis of the data available to us in order to make a structuring strategy for them and their extraction;
- We structured the data available to us and adjusted them to the needs of the model without compromising the consistency of the data;
- We developed a data extraction strategy. We chose to reduce the number of variables in order to check how much the model would have a built-in error;
- We decided to divide the software into two parts:
    - A part that will analyze and extract the data from the database in function of the parameter for elimination of the supplier depending on the total annual procurements from it;

- The second part will construct the regression model and based on the entered data it will predict what the price difference will be.

What We came to as a final conclusion is the following:

- The elimination of the number of variables was in order for the model to be more acceptable to the users and to eliminate the companies from which it was incidentally procured, is those that are insignificant for determining the dependent variable;
- Python is a good tool that can be used repeatedly to select different variants and evaluate the resulting model;
- It can be created a graphical interface for Python that would be very acceptable as an end user;
- Python can make a good service that can be placed on a server on the Internet or on a modern Cloud platform and be used by users as an on-demand concept and the results can be embedded in end-user interfaces;
- Having in mind the simple way of repetitive use, the models made in this way (with dynamic data extraction) can be used in the processes of BI and machine learning.
-

REFERENCES

Damadar. N.Gujarati, (2004) *BasicEconometrics*, FourthEdition, Tata McGraw Hill.

E. Malinvaud, (1966) *Statistical Methods of Econometrics*, Rand McNally, Chicago.

H. Theil *(*1971*), Principles of Econometrics*, John Wiley & Sons, New York.

Liang, S.,Li,X., Wang, J.,(2012). *Advanced Remote Sensing: Terrestrial Information Extraction and Applications, Academic Press*, pp. 800.

Spanos. A. Spanos (1999), *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, United Kingdom, 1999.

Zuur, A. K., Ieno, E.N., Smith, G. M., 2007. *Statistics for Biology and Health: Analyzing Ecological Data*, Springer Science + Business Media, LLC.

Taha, Hamdy A (2007), *OperationsResearch:An Introduction*8th ed., Pearson Prentice Hall.