DEPARTAMENTO DE ENGENHARIA INFORMÁTICA

MATCP Report

June 18, 2022

1DI, Group 40

1191296, Gabriel Gonçalves

1191369, Tiago Leite

1211314, João Durães

1211304, Francisco Bogalho

1210805, António Bernardo

Instituto Superior de Engenharia do Porto

Sandra Luna (SLU)

Index

1. Simple Linear Regression

Overview of Simple Linear Regression

Simple linear regression is a methodology developed from statistics and econometrics. This method uses a single explanatory variable in order to describe and estimate the relationship between two quantitative variables, that is, one dependent on the other (Y depends on X). Through this relationship we were able to have a graph, with great precision, predicting the values of the dependent variable with the values of the independent variable.

To create this graph, most of the time it is necessary to use a method called ordinary least squares, the objective of which is to minimize the sum of these squares of deviations, as much as possible.

With this graph we can calculate the correlation coefficient, this coefficient allows us to see how much data is needed. As a rule, we look for a correlation coefficient greater than 0,90.

Here, we can see the regression line equation:

$$Yi = \hat{a} + \hat{b}x + \varepsilon\mathrm{i}$$

Simple Linear Regression Model

In this project, an .xlsx file was used with all the necessary data, such as new cases, new deaths, reproduction rate, ICU patients, hospital patients, new exams, positive rate and fully vaccinated people). Since we have new cases and new deaths as dependents, keep in mind that all the others are independent.

The objective was to make a daily and weekly analysis and to study all possible relationships.

Regarding the Simple Linear Regression Model, there are 12 different relationships, so for the purposes of this user manual, we will only show the most relevant ones.

Model significance

This model is divided into a daily and weekly analysis and there are 12 different relationships, which means that there are 24 different Anova tables of correlation coefficients, confidence intervals and hypothesis tests.

After doing the necessary math, we were able to conclude that of these 24 values only a few are significant. That is, if in fact the largest of them present correlation coefficients.

After analyzing the Anova table, we draw a conclusion with the value 713,730 (Fo). Therefore, we conclude that we can admit that a given regression is linear if the value of Fo is greater than $f_{\alpha;1;n-2}$ (6,85).

Regarding the most significant models were the Y1-X5 relationship, new cases with a positive rate, and Y2-X5, new deaths with a positive rate, however, there is an overview of the entire study with regard to Simple Linear Regression.

Daily		
Relationship	**Regression line**	**Correlation coefficient**
new_cases-> reproduction_rate	**-2824,5x + 3422,7**	**0,1603**
new_deaths-> reproduction_rate	**-174,82x + 183,39**	**0,3347**
new_cases-> icu_patients	**4,3399x - 147,45**	**0,5986**
new_deaths-> icu_patients	**0,2199x - 23,914**	**0,8377**
new_cases-> hosp_patients	**0,6896x + 15,108**	**0,7326**
new_deaths-> hosp_patients	**0,0336x - 13,628**	**0,9484**
new_cases-> new_tests	**0,0049x + 885,97**	**0,0033**
new_deaths-> new_tests	**-0,0002x + 44,629**	**0,0024**
new_cases-> positive_rate	**30986x - 37,059**	**0,8581**
new_deaths-> positive_rate	**1421,9x - 13,01**	**0,9847**
new_cases-> people_fully_vaccina	**-0,0014x + 2031,5**	**0,1999**
new_death-> people_fully_vaccina	**-0,00007x + 89,262**	**0,3122**

Weekly		
Relationship	Regression line	Correlation coefficient
new_cases-> reproduction_rate	-2945,1x + 24642	0,1853
new_deaths-> reproduction_rate	-181,55x + 1321,5	0,3568
new_cases-> icu_patients	4,377x - 1126,7	0,6643
new_deaths-> icu_patients	0,2204x - 169,25	0,8535
new_cases-> hosp_patients	4,377x - 1126,7	0,6643
new_deaths-> hosp_patients	0,0337x - 96,414v	0,9601
new_cases-> new_tests	0,0049x + 885,97	0,0033
new_deaths-> new_tests	-0,0002x + 44,629	0,0024
new_cases-> positive_rate	31262x - 335,96	0,9392
new_deaths-> positive_rate	1430,4x - 93,144	0,996
new_cases-> people_fully_vaccina	-0,0015x + 14503	0,2277
new_death-> people_fully_vaccina	-0,00007x + 638,01	0,327

Daily

new_cases->positive_rate

Fonte variação	GL	Soma de quadratica	Média quadratica	F0
Regressão	1	245236314,792	245236314,8	713,7307351
Erro (residual)	118	40544541,133	343597,8062	
Total	119	1,23691E-10		

R^2	0,8581
R	0,9264

As the correlation coefficient table indicates, 0,858, is the closest to 1 and higher than 0,80, so that means it's a significative model and 85,81 % of the variation is explained by it.

new_deaths->positive_rate

Fonte variação	GL	Soma de quadratica	Média quadratica	F0
Regressão	1	516388,510	516388,5104	7573,108225
Erro (residual)	118	8046,081	68,18712938	
Total	119	9,37916E-12		

R^2	0,9847
R	0,9923

As the correlation coefficient table indicates, 0,984, is the closest to 1 and higher than 0,80, so that means it's a significative model and 98,47% of the variation is explained by it.

Weekly

new_cases->positive_rate

Fonte variação	GL	Soma de quadratica	Média quadratica	F0
Regressão	1	1699764595,411	1699764595	231,6242
Erro (residual)	15	110076873,648	7338458,243	
Total	16	2,91038E-11		

R^2	0,9392
R	0,9691

As the correlation coefficient table indicates, 0,939, is the closest to 1 and higher than 0,80, so that means it's a significative model and 93,92% of the variation is explained by it.

new_deaths->positive_rate

Fonte variação	GL	Soma de quadratica	Média quadratica	F0
Regressão	1	3558472,692	3558472,692	3711,1049
Erro (residual)	15	14383,072	958,8714917	
Total	16	1,25056E-12		

R^2	0,9960
R	0,9980

As the correlation coefficient table indicates, 0,996, is the closest to 1 and higher than 0,80, so that means it's a significative model and 99,60% of the variation is explained by it.

Hypothesis tests for model coefficients

The purpose of the Hypothesis Test is to analyze the relationship and decide on its results. We also test whether we can consider the parameter values equal to zero.

Here's what a hypothesis test looks like:

$$H0\colon \hat{a} = 0 \;\; v.s. \;\; H1\colon \hat{a} \neq 0 \qquad or \qquad H0\colon \hat{b} = 0 \;\; v.s. \;\; H1\colon \hat{b} \neq 0$$

Confidence intervals for prediction values

Depending on the confidence level, the purpose of these intervals is to calculate an interval to which we are sure the parameter value belongs.

To calculate a confidence interval, we need to specify a confidence interval, as requested - 90% and 95%, and determine the value of tc. Next, we calculate the standard deviation, which depends only on the number of samples and the values of the dependent variable. Finally, apply the formula and finally add and subtract everything with the corresponding parameter.

$$a \pm tcs\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}} \qquad b \pm tcs\sqrt{\frac{1}{Sxx}}$$

From this we can get the range - an upper and lower bound.

Daily				
----------------------------------	90%		95%	
----------------------------------	alpha	beta	alpha	beta
new_cases-> reproduction_rate	[2573,6; 4271,7]	[-3810,9; -1838,1]	[2408.5, 4436,8]	[-4002.8, -1646,2]
new_deaths-> reproduction_rate	[151,0; 215,8]	[-212,4; -137,2]	[144.7, 222.1]	[-219.8,-129.9]
new_cases-> icu_patients	[-29,7; -18,1]	[0,21; 0,23]	[-30.9,-17.0]	[0.03,0.04]
new_deaths-> icu_patients	[-140,8; 171,0]	[0,63; 0.75]	[-171.2,201,4]	[0.61,0.77]
new_cases-> hosp_patients	[-16,6; 10,7]	[0,03; 0,03]	[-17.1,-10.1]	[0.03, 0.04]
new_deaths-> hosp_patients	[341,6; 1430,3]	[-0,008; 0,018]	[235.8,1536.2]	[-0.012, 0.021
new_cases-> new_tests	[21,3; 67,9]	[-0,0007; 0,0004]	[16.8, 72.5]	[-0.0009,0.0005]
new_deaths-> new_tests	[-149,3; 75,2]	[29063,4; 32909,2]	[-171.1, 97,03]	[28689.5, 33283.1]
new_cases-> positive_rate	[-14,6; -11,4]	[1394,8; 1448,9]	[-14.9, -11.1]	[1389.5,1454.2]

	90% alpha	90% beta	95% alpha	95% beta
new_deaths-> positive_rate	[1370,5; 2392,5]	[-0,002; -0,001]	[1600.3, 2462.7]	[-0.002,-0.0009]
new_cases-> people_fully_vaccina	[74,9; 103,6]	[-0,0001; -0,0001]	[72.13,106.4]	[-0.0001,-0.0001]
new_death-> people_fully_vaccina	[-360,7; 65,8]	[3,80; 4,88]	[-402.19, 107.3]	[3.69, 4.99]

Weekly				
----------------------------	90%		95%	
----------------------------	alpha	beta	alpha	beta
new_cases-> reproduction_rate	[7871,5; 41411,7]	[-5739,9; -150,2]	[4251,6; 45031,5]	[-6343,2; 453,1]
new_deaths-> reproduction_rate	[659,4; 1983,6]	[-291,9; -71,2]	[516,5; 2126,5]	[-315,7; -47,4]
new_cases-> icu_patients	[-5012,1; 2758,8]	[2,9; 5,8]	[-5850,8; 3597,5]	[2,6; 6,1]
new_deaths-> icu_patients	[-283,2988; -55,1929]	[0,1791 ; 0,2618]	[-307,9173; -30,5744]	[0,1702 ;0,2707]
new_cases-> hosp_patients	[-2585,8386 ;2588,8102]	[0,5473 ; 0,8488]	[-3144,3174 ;3147,2890]	[0.51,0.88]
new_deaths-> hosp_patients	[-149,7420 ; -43,0860]	[0,0306 ; 0,0368]	[-161,3; -31.6]	[0,03; 0,04]
new_cases-> new_tests	[341,6114; 1430,3289]	[-0,0081; 0,0180]	[-10703; 25396,1]	[-0,06; 0,06]
new_deaths-> new_tests	[21,2988; 67,9586]	[-0,0007; 0,0004]	[-373,4; 1220,2]	[-0,004; 0,002]
new_cases-> positive_rate	[-1801,9699; 1130,0403]	[27660,846; 34862,6797]	[-2118,4; 1446.5]	[26883,5; 35639,9]
new_deaths-> positive_rate	[-109,9016;-76,3863]	[1389,2156; 1471,5390]	[-113,5; 72,8]	[1380,3; 1480,4]
new_cases-> people_fully_vaccina	[7392,7282; 21614,2330]	[-0,0027;-0,0002]	[5857,9; 23149,1]	[-0,003; 0]
new_deaths-> people_fully_vaccina			[279,4; 996,6]	[-0,0001; -0,0001]

Multiple Linear Regression

Overview of Multiple Linear Regression

As its name implies, Multiple Linear Regression is also a linear regression model, although it uses multiple explanatory variables, as opposed to Simple Linear Regression.

The purpose of these regressions is to study the relationship between these variables - one depends on many other independent ones.

Bearing in mind that the MLR works with more than two variables, there is no such thing as a regression line like the SLR. Then, we calculate the correlation coefficients for each variable.

The regression model looks like this:

$$yi = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \cdots + \beta kxk + \varepsilon i$$

To calculate these coefficients, we first need to calculate some matrices:

- **The X matrix** – where the first column is formed by 1's and the others columns are composed by the values of the independent variables.
- The X matrix **transposed**
- X matrix transposed **times** the X matrix
- The **inverse** of the matrix above.
- Finally, the X matrix transposed **times** the dependent variable values.

Then, by determining these matrices, we just have to multiple the inverse matrix with the last one mentioned. By calculating these, we will have a different correlation coefficients.

After calculating the coefficients, we can estimate the values of the dependent variables with the values of the independent variables.

Like the SLR, in the MLR, to better explore the relationship between these variables, these coefficients were also estimated with hypothesis tests and V.

Finally, the Anova table is also used to take decisions about the results.

Multiple Linear Regression Model

Similar to the SLR analysis, this one is also divided into daily and weekly analysis, however, unlike the SLR, instead of having twelve different relationships, here we have only two, which are the new cases and new deaths will be all the others.

As there are two dependent and six independent variables, then there are seven correlation coefficients in each relationship.

For each coefficient, there are confidence intervals and hypothesis tests.

In addition, there is an Anova table, which helps in decision making.

Due to the user manual, we will only show and discuss the results achieved.

Model significance

By looking at the Anova table, we can decide based on the value of Fo. We can conclude that it is acceptable to admit that a given regression is linear if the value of Fo is greater than fα;k;n-(k+1).

After the calculations, we came to the conclusion that the 2 relationships, both in the daily and weekly analysis, present high coefficients of determination, which means that they explain the variance of the data well.

The results obtained can be found on the next page (Table Anova and Coefficient of Determination).

Hypothesis tests for model coefficients

As mentioned earlier in Simple Linear Regression, hypothesis tests serve to conclude whether we can consider the parameters equal to zero. In this case, in Multiple Linear Regression the objective is the same, instead of the parameters, we test all the coefficients.

Here's what a hypothesis test looks like

$$\mathrm{H}_0: \beta_j = 0 \;\; v.s \;\; H_1: \beta_j \neq 0$$

In order to make a decision we have to check whether or not To is higher than tc

$$t_o = \frac{\beta}{\sqrt{\delta^2 C_{jj}}}$$

So, if To is higher than tc, we reject Ho.

Here we have an overview of the entire study on Multiple Linear Regression

Confidence intervals for prediction values

The purpose of Confidence Intervals is to calculate an interval depending on a given confidence level.

At intervals of 100, the value of certain coefficients is within 90 or 95 of these intervals.

To calculate and confidence interval first we need to select the coefficient. Then, we need to calculate the standard deviation and use the corresponding value of the C_{jj} value.

$$\beta_j \pm \sqrt{\delta^2 C_{jj}}$$