FINDING A GENERAL FORMULA FOR DATA PREDICTION IN LINEAR REGRESSIONS

CANDIDATE NAME :SIMONE ZAMBETTI

CANDIDATE NUMBER: 000815-0123

SCHOOL: HOCKERILL ANGLO-EUROPEAN COLLEGE

YEAR OF SUBMISSION: MAY 2014

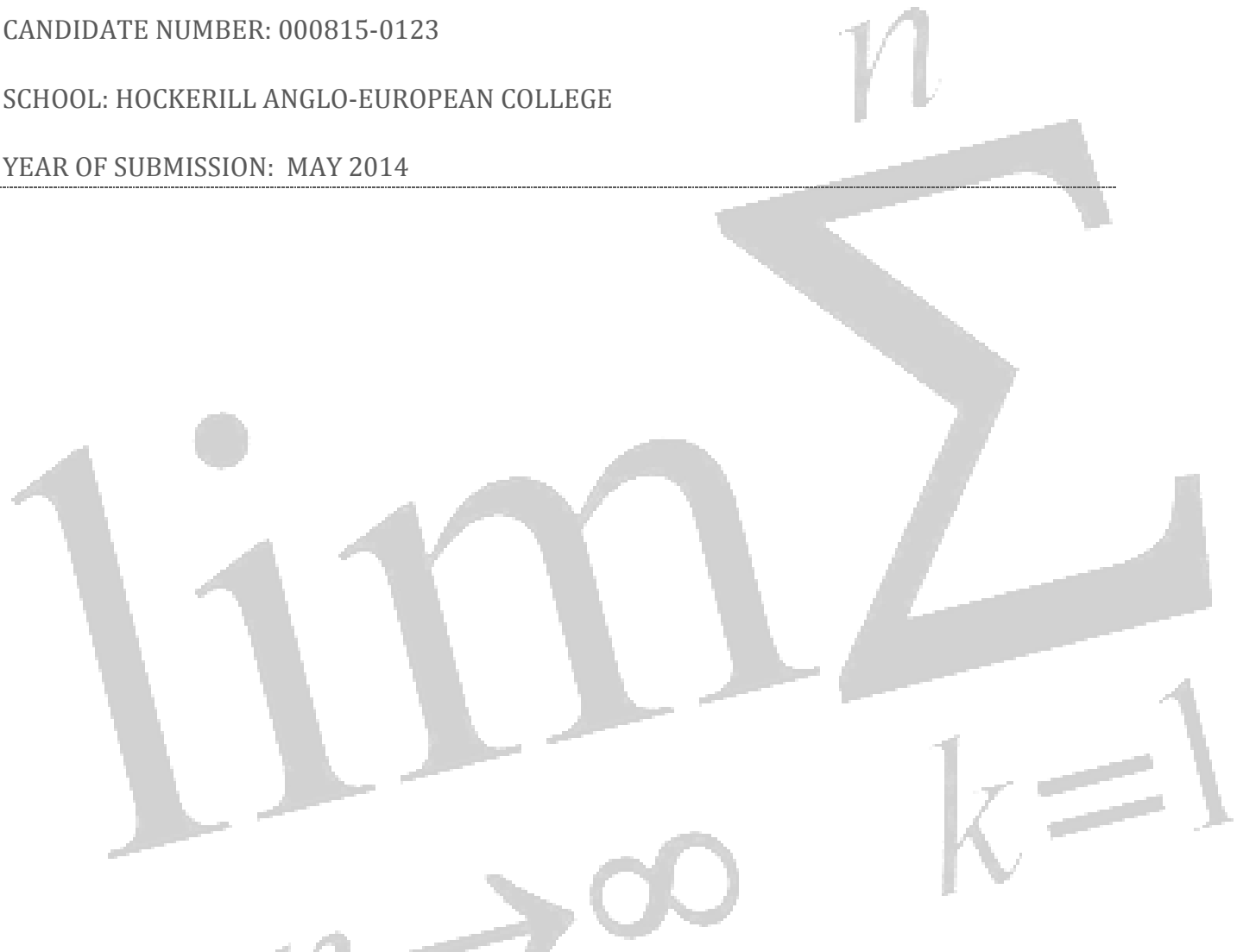


Table of contents

RATIONALE AND INTRODUCTION

Taking Mathematics, Physics and Biology during my IB diploma, I had to graph many so-called "best fitting lines" using various software or even by hand, in order to demonstrate a relation between two random variables. The scope of such a procedure is to draw a line which best represents the trend followed by data, thus enabling the prediction of its future location. But how do these softwares graph these lines? – I wondered each time. Logically, I came to the conclusion that there must be some general rule to that they follow. From this, the curiosity of finding this universal formula to plot best fitting lines.

Now, it is very easy to perform such a procedure when data perfectly fits in a line. However, in a reality this is never the case, as the data collected will present variations due to the random factors of one or both variables being studied.

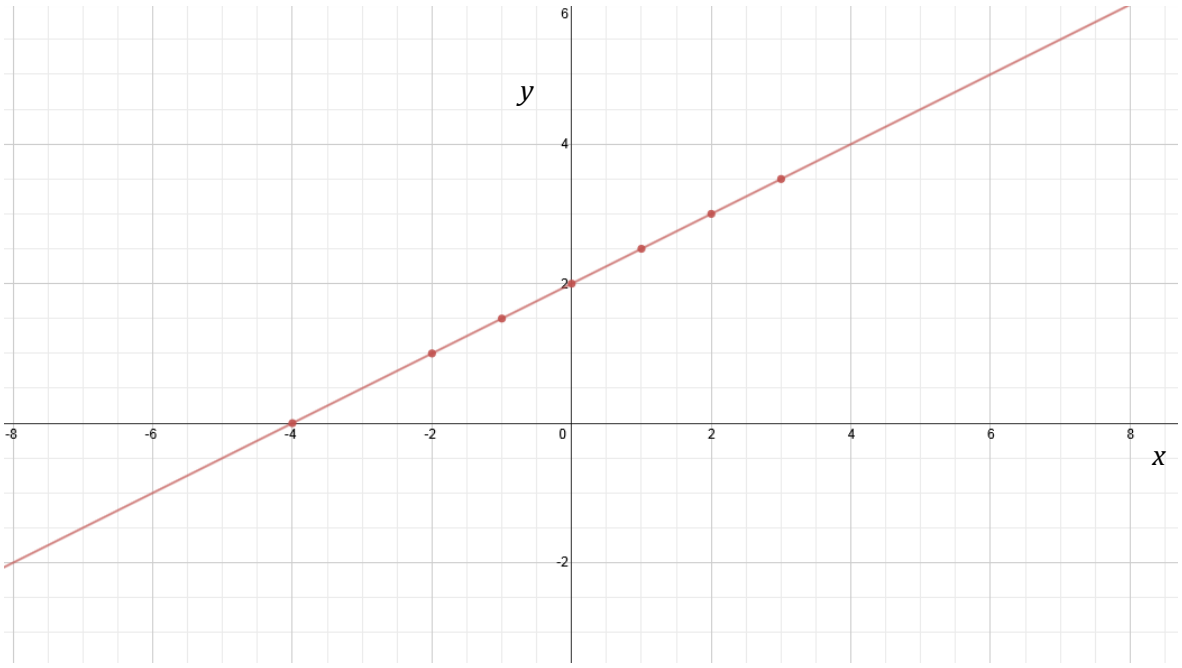


Figure 1– Linear Correlation

If data (red dots on figure 1) is aligned, the line of best fit will intersect every single value. To find its equation it is enough applying the very basic formula:

$$y - y_1 = m\,(x - x_1) \qquad [1]$$

Therefore, by calculating the equation of the line using (0,2) and (-4,0), the result is:

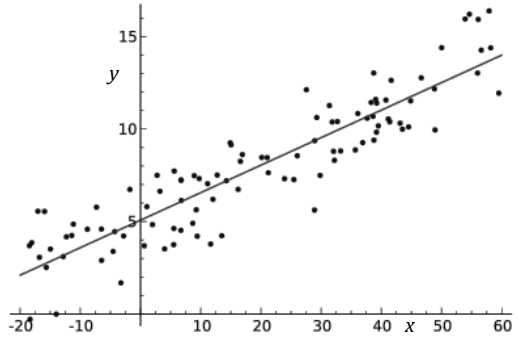
$$y = \frac{1}{2}\,x + 2 \qquad [2]$$

With these very basic Math tools, it has just been confirmed a pattern correlating an event *X* with an event ***Y***, which has been found to be dependent to *X* itself.

From such, another very important idea derives: what if the data points were not perfectly aligned? How would we individuate a best fit line? To explain adequately such a concept, we will be forced to introduce multiple notions, which will be tackled in the following chapters.

LINEAR REGRESSION

In statistics, linear regression is one of the approaches to model the relationship between a scalar dependent variable Y and one explanatory[1] variable, the so-called X. The case of one explanatory variable is called linear regression. In this model, data is modeled using linear predictor functions, and unknown model parameters are estimated from the data. Linear regression refers to a model in which the conditional mean of Y given the value of X is an dependent function of X, or rather is related on X itself.



This graph represents points of regression of a random variable *X*, along with the best fit line of data. But how do we know the equation of the line? How do softwares like Excel provide us with a specific equation?

Figure 2 – A graph with an explanatory variable *x*

[1] Same as independent

I thought it would have been easier to simply things as much as possible. Therefore we start off with a regression graph where there is a number n of points. This can be visualized with the following graph:

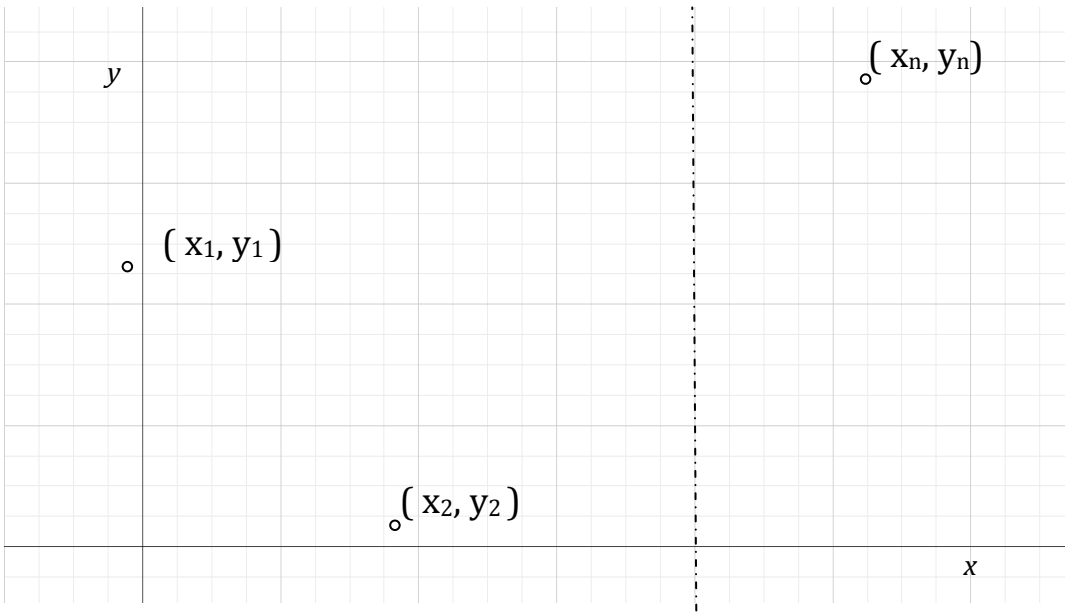


Figure 3 – n points and their coordinates on a Cartesian plane. The pointed line stands for a virtual cut in the plane itself.

For the sake of simplicity, I assume that, despite there would be n points on the plane (cut out by the dotted line), just three points are visible on the graph: their generic coordinates are shown.

As stated, what we want to do is find the equation of a line so that it best represents the trend followed by the previous points. Illustrating:

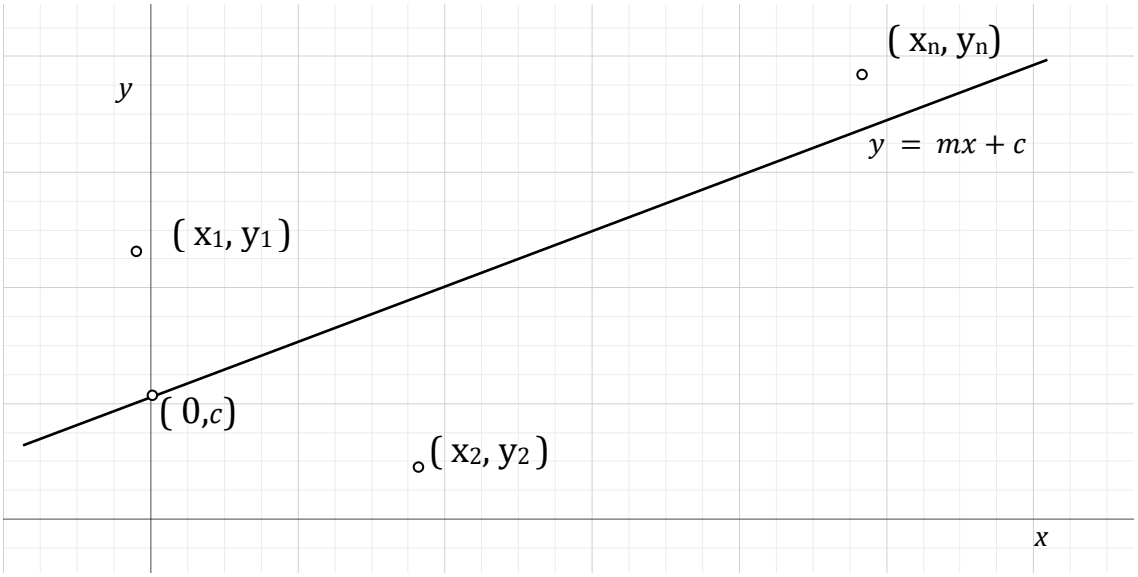


Figure 4 – Drawing the alleged best fit line in a n points graph

The line graphed is the alleged best fit line of the system. Being a line, it will possess an equation in the form of $y = mx + c$, where m is the coefficient and c the intercept on the y axis, obtained by setting $x = 0$.As the line must represent the data with the smallest possible error, this means that the overall distance from the data to the line must be the smallest possible. Hence, the position of the best fit line must be given by the smallest sum of the squared distances (in order to avoid the direction of the error, so that it is always positive) from the data points themselves and the best fit line. This is commonly called the *Least Squares* procedure.

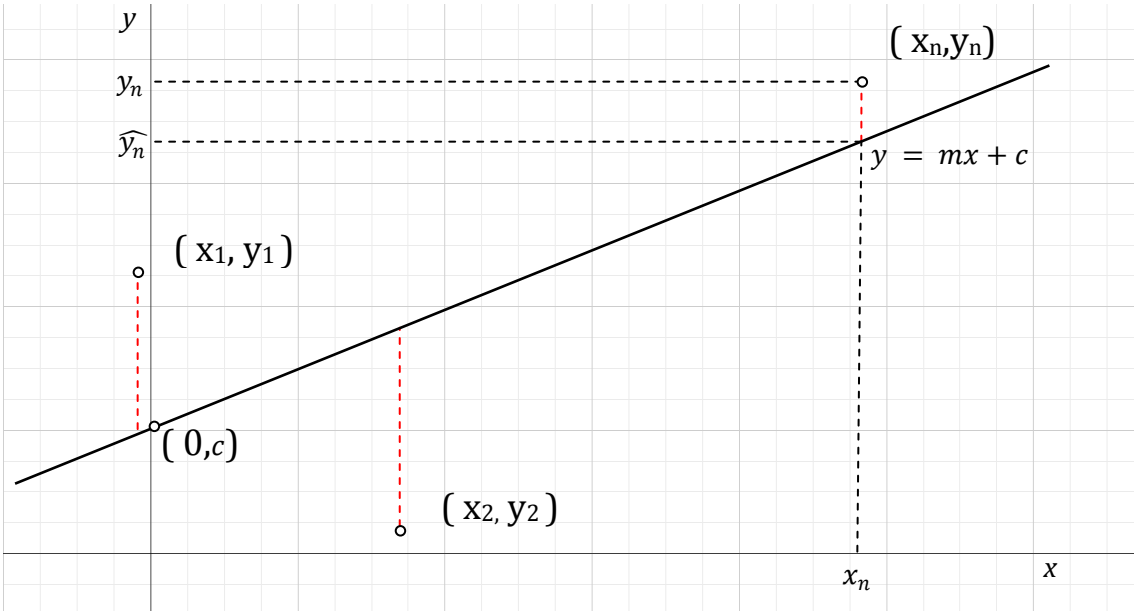


Figure 5 - the distances from the points and the axis y and x

The aim of the previous graph is to indicate errors between the best fit line and the actual data points. An example of this is e_n, which is equal to the actual y value of the n[th] point y_n minus the corresponding value on the line of best fit, $\widehat{y_n}$.

$$e_n = y_n - \widehat{y_n} \qquad [3]$$

As $\widehat{y_n}$ lies on the line of best fit, it can be rearranged in the form:

$$e_n = y_n - (m\,x_n + c) \qquad [4]$$

We therefore have just found a formula to calculate the distance actual point – predicted point that applies to all the data in the graph. Generalizing the formula for the *n* data points (y_i, x_i), where $i = 1, 2, \ldots, n$. :

$$e_i = y_i - \widehat{y_i} = y_i - (mx_i + c) \qquad [5]$$

Now, the e_i values found are squared and summed. Visualizing this passage on the graph will help comprehend the aims of such a procedure.

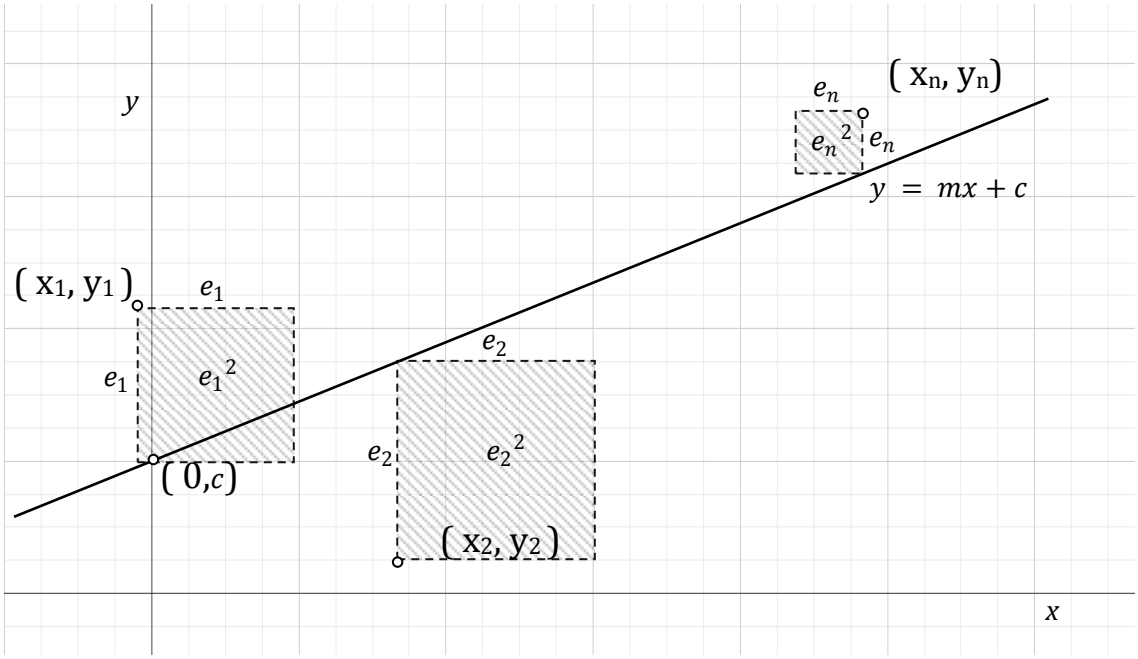


Figure 6 - The squares generated by squaring the errors from the line

By squaring the distances e_i, we formed n squares, having as sides the distance of each point from the line (e_1 , e_2 , e_n). Through this method, negative and positive distances do not need to be taken into consideration as it will be considered the sum of the areas, which is always positive.

Recalling equation number 5, the sum of the squares of the e_i can be therefore be expressed as:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \equiv \sum_{i=1}^{n} [\, y_i - (m\, x_i + c)]^2 \qquad [\,6\,]$$

Expanding the previous:

$$\sum_{i=1}^{n} [\, y_i - (m\, x_i + c)]^2 = \sum_{i=1}^{n} y_i^2 - 2y_i(m\, x_i + c) + (m\, x_i + c)^2 \qquad [\,7\,]$$

Further expansion from 7, aiming to simplify further:

$$\sum_{i=1}^{n} y_i^2 - 2y_i m\, x_i - 2y_i c + m^2 x_i^2 + 2cmx_i + c^2 \qquad [\,8\,]$$

Thus, this formula provides us the sum of the areas of the distances between the best fit line and the data points. However, I thought the sigma notation is not very useful to visualize what concerns the following passages. Therefore, we need to manipulate algebraically again the previous equation into the following:

$$\sum_{i=1}^{n} e_i^2 = (y_1^2 + y_2^2 + y_3^2 \ldots + y_n^2) + 2m\,(x_1 y_1 + x_2 y_2 + \ldots x_n y_n) - 2c(y_1 + y_2 + \ldots y_n) + m^2(x_1^2 + x_2^2 + x_3^2 \ldots + x_n^2) + 2mc\,(x_1 + x_2 + \ldots x_n) + nc^2 \qquad [\,9\,]$$

However it is true that:

$$\sum_{i=1}^{n} x_i y_i \frac{1}{n} \equiv \frac{y_1^2 + y_2^2 + y_3^2 \ldots + y_n^2}{n} = \overline{xy} \qquad [\,10\,]$$

And therefore:

$$\mathrm{y}_1^2 + \mathrm{y}_2^2 + \mathrm{y}_3^2 + \ldots + \mathrm{y}_\mathrm{n}^2 += \overline{\mathrm{y}^2}\mathrm{n} \qquad [\,11\,]$$

Similarly, the same process can used for the other variables within the equation 9:

$$\sum_{i=1}^{n} x_i y_i \frac{1}{n} \equiv \frac{x_1 y_1 + x_2 y_2 + \ldots x_n y_n}{n} = \overline{xy} \quad [12]$$

$$\therefore\ x_1 y_1 + x_2 y_2 + \ldots x_n y_n = \overline{xy}\, n \quad [13]$$

$$\sum_{i=1}^{n} y_i \frac{1}{n} \equiv \frac{y_1 + y_2 + \ldots + y_n}{n} = \bar{y} \quad [14]$$

$$\therefore\ y_1 + y_2 + \ldots y_n = \bar{y}\, n \quad [15]$$

$$\sum_{i=1}^{n} x_i \frac{1}{n} \equiv \frac{x_1 + x_2 + \ldots + x_n}{n} = \bar{x} \quad [16]$$

$$\therefore x_1 + x_2 + \ldots + x_n = \bar{x} n \quad [17]$$

$$\sum_{i=1}^{n} x_i^2 \frac{1}{n} \equiv \frac{x_1^2 + x_2^2 + \ldots + x_3^2}{n} = \overline{x^2} \quad [18]$$

$$\therefore\ x_1^2 + x_2^2 + x_3^2 \ldots . + x_n^2 = \overline{x^2} n \quad [19]$$

Now we can use 11, 13, 15, 17, 19 to substitute them into equation 9: the aim of such is to have the sum of the squares the most as simplified as possible, as a further manipulation is needed.
It is therefore obtained the equation for the squared distances, also called as SD:

$$SD = \sum_{i=1}^{n} e_i^2 \equiv \overline{y^2} n + 2m\, \overline{xy}\, n - 2c\, \bar{y}\, n + m^2 \overline{x^2} n + 2mc\, \bar{x} n + nc^2 \quad [20]$$

Thus we now want to optimize the formula obtained for the sum of the distances squared in 19, in order to find its minimum. The target of this step is that if the total area of the squares represented in figure 6 is the lowest possible, then also the overall distance from the data points to the best fit line will be the smallest possible, which results in the best estimation factor, as the line will be the closest possible to the all the actual data, along with the best overall representation.

To minimize the error, it is necessary to partially derive the equation obtained in 20.

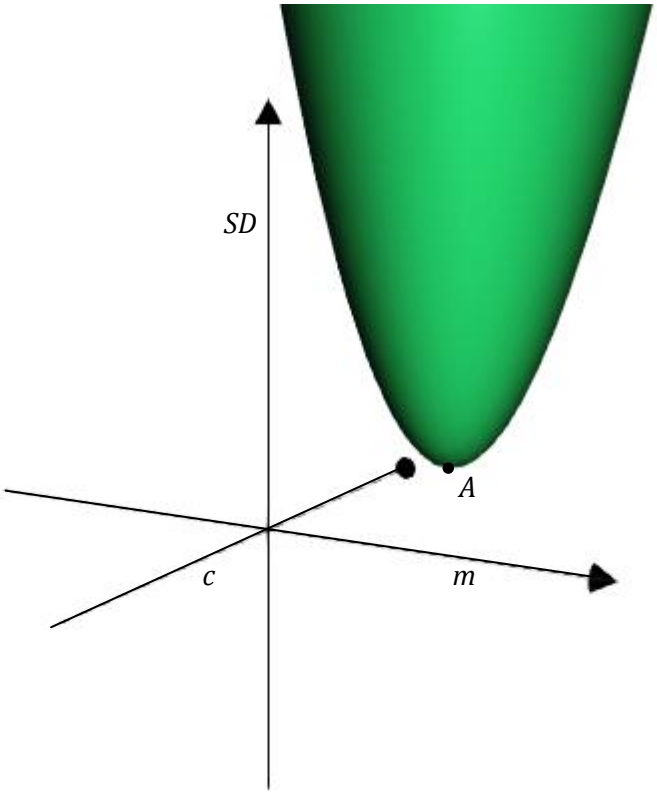


Figure 7 - The 3D equation when SD is plotted and its minimum, point A.

As the line SD has an equation in the form of $y = mx + c$, with m and c as unknowns, we want to find their value so that the errors areas are the smallest. Due to this necessity, a 3D graph has been plotted. The axis have been parted into m, c and SD function. Point A indicates the minimum magnitude of the sum of the errors squared. At such point, the following can be stated, as it is a minimum of the SD function:

$$\frac{d}{dm}SD = 0 \ \ and \ \ \frac{d}{dc}SD = 0 \qquad [\ 21\ ]$$

Finding the partial derivatives of 19, with respect of m and c:

$$\frac{d}{dm}SD = \ -2n\overline{xy} + \ 2nm\overline{x^2} + \ 2cn\bar{x} \qquad [\ 22\ ]$$

$$\frac{d}{dc}SD = -2n\bar{y} + 2nm\bar{x} + 2cn \qquad [\ 23\ ]$$

Therefore, as we want to find point A, in order to obtain the smallest c, m and the smallest sum of squares, $\frac{d}{dm} SD$ and $\frac{d}{dc} SD$ must be equal to 0. What follows are 22 and 23 further simplified:

$$2n\,(-\overline{xy} + m\overline{x^2} + c\bar{x}) = 0 \qquad [\,24\,]$$

$$2n\,(-\bar{y} + m\bar{x} + c) = 0 \qquad [\,25\,]$$

As 2n is never equal to 0, as the number of data points cannot be 0, it is obvious that:

$$given\ that\ 2n \neq 0 \ \Rightarrow\ -\overline{xy} + m\overline{x^2} + c\bar{x} = 0 \ \ and \ \ -\bar{y} + m\bar{x} + c = 0 \qquad [\,26\,]$$

From this I noticed that the equation in 26 can be rearranged even further to finally extrapolate the value of the linear coefficient, m and the y-intercept c.

Rearranging from equation 26:

a) $\overline{xy} = m\overline{x^2} + c\bar{x}$,manipulating even further: $\frac{\overline{xy}}{\bar{x}} = m\frac{\overline{x^2}}{\bar{x}} + c$

b) $\bar{y} = m\bar{x} + c$

What has just been found by manipulating these expressions is very notable. In fact, implicitly it has just been proven that the points $(\bar{x}\,, \bar{y})$ and $\left(\frac{\overline{x^2}}{\bar{x}}\,, \frac{\overline{xy}}{\bar{x}}\right)$ are on the regression best fit line from the equations a) and b).

Now it is therefore possible to calculate the linear coefficient of the best fit line, which has been called as m during the developing of the study. Rearranging the formula of the line through the equation in 1, it is obtained that it is:

$$m\left(x - \frac{\overline{x^2}}{\bar{x}}\right) = y - \frac{\overline{xy}}{\bar{x}} \qquad [\,27\,]$$

Thus the m can be obtained using the coordinates of the points found previously through the expression:

$$m = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} \qquad [\,28\,]$$

It is now obvious that 27 and 28 can be rearranged together, in order to obtain the best fit line equation:

$$y - \frac{\overline{xy}}{\bar{x}} = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}}\left(x - \frac{\overline{x^2}}{\bar{x}}\right) \quad [29]$$

Rearranging 29:

$$y = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}}x - \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} \times \frac{\overline{x^2}}{\bar{x}} + \frac{\overline{xy}}{\bar{x}} \quad [30]$$

Simplifying further 30:

$$y = \frac{\frac{\bar{y}\bar{x} - \overline{xy}}{\bar{x}}}{\frac{(\bar{x})^2 - \overline{x^2}}{\bar{x}}}x - \frac{\frac{\bar{y}\bar{x} - \overline{xy}}{\bar{x}}}{\frac{(\bar{x})^2 - \overline{x^2}}{\bar{x}}} \times \frac{\overline{x^2}}{\bar{x}} + \frac{\overline{xy}}{\bar{x}} \quad [31]$$

$$y = \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}x - \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} \times \frac{\overline{x^2}}{\bar{x}} + \frac{\overline{xy}}{\bar{x}} \quad [32]$$

$$y = \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}x - \frac{\overline{x^2}(\bar{y}\bar{x} - \overline{xy})}{(\bar{x})^3 - \bar{x}\overline{x^2}} + \frac{\overline{xy}}{\bar{x}} \quad [33]$$

$$y = \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}x + \frac{\overline{x^2}\bar{y}\bar{x} - \overline{x^2}\overline{xy}}{-(\bar{x})^3 + \bar{x}\overline{x^2}} + \frac{\overline{xy}}{\bar{x}} \quad [34]$$

$$y = \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}x - \frac{\overline{x^2}\bar{y}\bar{x} - \overline{x^2}\overline{xy} - (\bar{x})^2\overline{xy} + \overline{x^2}\overline{xy}}{(\bar{x})^3 - \bar{x}\overline{x^2}} \quad [35]$$

$$y = \frac{\bar{y}\bar{x} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}}x - \frac{\overline{x^2}\bar{y}\bar{x} - (\bar{x})^2\overline{xy}}{(\bar{x})^3 - \bar{x}\overline{x^2}} \quad [36]$$

Re-expanding 36, using the results of 12, 13, 14,15, 16, 17, 18 and 19:

$$y = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i}{\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2} x + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i y_i \cdot \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^3 - \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2} \quad [\,37\,]$$

$$y = \frac{\frac{1}{n^2}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i}{\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right]^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2} x + \frac{\frac{1}{n^3}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \frac{1}{n^3}\sum_{i=1}^{n} x_i y_i \cdot \left(\sum_{i=1}^{n} x_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^3 - \frac{1}{n}\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\sum_{i=1}^{n} x_i^2} \quad [\,38\,]$$

$$y = \frac{\frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\, y_i\right)}{\frac{1}{n}\left\{\frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2 - \sum_{i=1}^{n} x_i^2\right\}} x + \frac{\frac{1}{n^3}\left(\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \cdot \left(\sum_{i=1}^{n} x_i\right)^2\right)}{\frac{1}{n^2}\left\{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^3 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i^2\right\}} \quad [\,39\,]$$

$$y = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\, y_i}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2} x + \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \cdot \left(\sum_{i=1}^{n} x_i\right)^2\right)}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^3 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i^2} \quad [\,40\,]$$

$$f(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\, y_i}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2} x + \frac{\frac{1}{n}\left(\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \cdot \sum_{i=1}^{n} x_i\right)}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2} \quad [\,41a\,]$$

$$\text{and } f(x) \text{ exists } \forall x \text{ where } \mathrm{D}x\colon \{x \in \mathbb{R}\} \Leftrightarrow (x/(;Im) = 0), \quad \left\{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2\right\} \neq 0 \quad [\,41b\,]$$

We have just found the general formulae of a best fitting line through n data points, where n is the number of values given, x_i is the x coordinate of a data point for which $i = 0, 1, 2, 3 \ldots n$, and similarly y_i is the y coordinate of any data give, where $i = 0, 1, 2, 3 \ldots n$.

USING THE FORMULAE

Inthis last section, it is shown that the formula actually works and provides the same result of excel. Three points were decided randomly, and they resulted to be:

	x values	y values
A	2	6
B	4	7
C	5	3

Table 1 – Sample Data points coordinates

Using these points, excel plots the line providing the following equation:

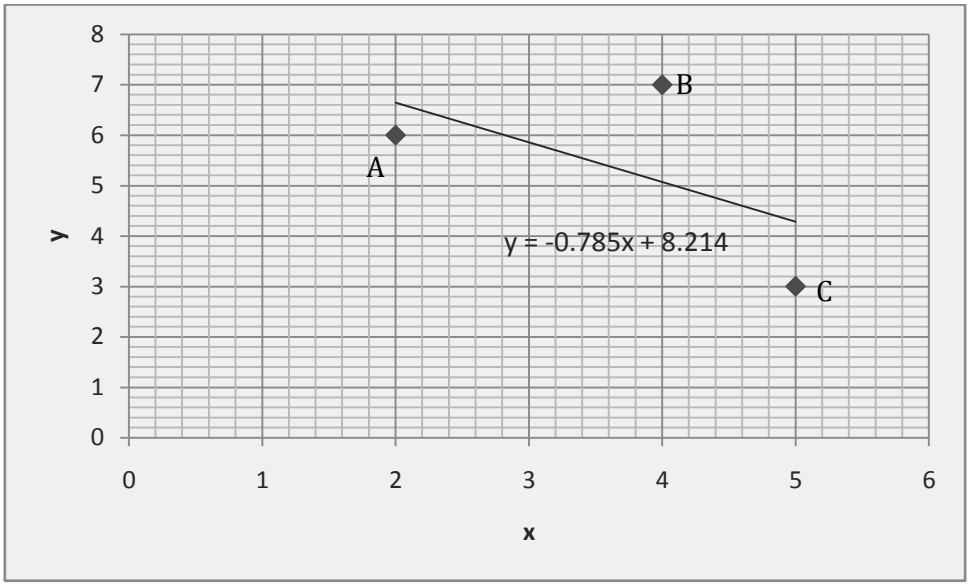


Figure 8 – Excel best fit line equation through the three points

Therefore using the formula in 41a:

$$f(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\, y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2} x + \frac{\frac{1}{n}\left(\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i \cdot \sum_{i=1}^{n} x_i\right)}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2}$$

$$\therefore m = \frac{\frac{1}{3}(2+4+5)(6+7+3) - (12+28+15)}{\frac{1}{3}(2+4+5)^2 - (4+16+25)} = -0.785$$

$$\therefore c = -\frac{\frac{1}{3}[(6+7+3)(2^2+4^2+5^2) - (12+28+15)(2+4+5)]}{\frac{1}{3}(2+4+5)^2 - (4+16+25)} = 8.214$$

$\rightarrow f(x) = -0.785x + 8.124$, which is the same equation as calculated by Excel in Figure 8.

Thus the formula in 42a is valid; the method and procedure followed in the exploration are correct and have been confirmed.

CONCLUSION

Finding best fitting lines in order to forecast data is essential in many areas: in modern econometrics for instance, despite other statistical tools are frequently used, linear regression is still the most frequent starting point for an analysis. Consider Okun's law: it states that percentage GDP growth is correlated with the change in unemployment rate. This relationship is represented in a linear regression where the change in unemployment rate is a function of an intercept, a given value of GDP growth multiplied by a slope coefficient m.

The equation to describe the linear regression correlation has been found, and resultsare correct. I therefore have calculated a general formula used by software like Excel to plot a linear best fit line using non-aligned data points. The scope of the exploration has been therefore achieved successfully.

The procedure itself has been an output of two months of researching: I had the idea of using the least squares when I saw a friend of mine modeling a cubic, for example. However, the final formula is quite complex, having a number of factors, thus raising the probability of making mistakes during calculations. For instance, when I was rearranging or typing the values on the calculator to make sure the steps were consistent, it was almost never right the first time, thus forcing its checking multiple times.

The formula exists for every set of values in the Cartesian plane: this is a significant advantage, as the equation will be provide consistent results whatever the data values are assumed to be, provided to operate in the real set of numbers.

As the exploration itself arises from personal curiosity, I consequently have been enthusiastic whilst carrying out the entire exploration. This is true particularly in the final part of the exploration itself, when I realized my equation was correct, showing the same values of Microsoft Excel.

I think I could have explored regression patterns even more, for example introducing briefly a general equation for polynomials; however, the length and complexity of the content, such as graphs in 4 dimensions, would not have allowed a complete explanation.In addition, the example given may have been better if related to econometrics, perhaps the Okun’s law.

BIBLIOGRAPHY:

Simple linear regression - Wikipedia, the free encyclopedia. 2014. SIMPLE LINEAR REGRESSION - WIKIPEDIA, THE FREE ENCYCLOPEDIA. [ONLINE] Available at: http://en.wikipedia.org/wiki/Simple_linear_regression. [Accessed 12 August 2013].

Rensselaer Polytechnic institute. 2012. . [ONLINE] Available at: http://homepages.rpi.edu/~tealj2/stat02.pdf. [Accessed 17 August 2013].

Econometrics - Wikipedia, the free encyclopedia. 2014. ECONOMETRICS - WIKIPEDIA, THE FREE ENCYCLOPEDIA. [ONLINE] Available at: http://en.wikipedia.org/wiki/Econometrics. [Accessed 22 August 2013].

JSTOR: .1974. JOURNAL OF THE ROYAL STATISTICAL SOCIETY. SERIES C (APPLIED STATISTICS)Vol. 23, No. 3 [ONLINE] Available at: http://www.jstor.org/discover/10.2307/2347147?uid=3738032&uid=2&uid=4&sid=21103350532073. [Accessed 13 September 2013].

P.JoanneCornbleet and Nathan Gochman .1979.INCORRECT LEAST-SQUARES REGRESSION COEFFICIENTS IN METHOD- COMPARISON ANALYSIS [ONLINE] AVAILABLE AT: http://www.clinchem.org/content/25/3/432.long[Accessed 1 October 2013].

Regression and Correlation Methods in StatsDirect. 2014. REGRESSION AND CORRELATION METHODS IN STATSDIRECT. [ONLINE] Available at: http://www.statsdirect.com/help/content/regression_and_correlation/regression_and_correlation.htm. [Accessed 3 October 2013].

Graph plotter software [ONLINE] Available at http://www.livephysics.com/tools/mathematical-tools/online-3-d-function-grapher/