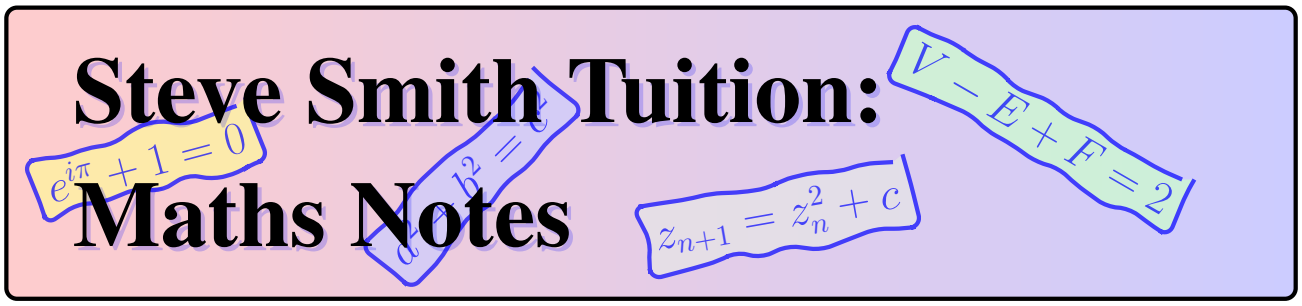

Finding Equations of Regression Lines

Contents

Prerequisites

I am assuming that you have covered the essentials of correlation and regression. In this document we are going to use what we know about correlation and regression to find the equation of the regression line (the best-fit straight line) through some data.

Notes

None.

Document History

Date	Version	Comments
8th December 2013	1.0	Initial creation of the document.

1 Introduction

Here we are just going to go over the very rudiments of correlation and regression, to remind ourselves what these terms mean.

1.1 A Quick Overview of Correlation

Two variables (x and y, say) are said to be correlated if there is some pattern, some relationship, between the values of the two variables in a given set of (x, y) data.

1.1.1 Positive Correlation

So for example, there is said to be *positive* correlation between the variables x and y if a set of (x, y) data looks like Figure 1.

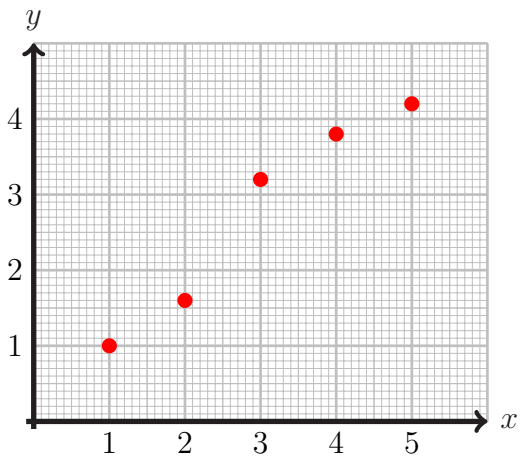

Figure 1: Positive Correlation

It's a positive correlation because, in general, y increases as x increases, so that if you were to draw a best-fit straight line through this data, it would have a positive gradient.

1.1.2 Negative Correlation

And there is said to be *negative* correlation between the variables x and y if a set of (x, y) data looks like Figure 2.

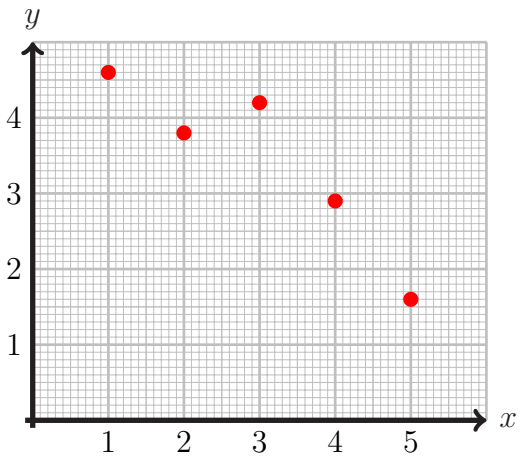

Figure 2: Negative Correlation

It's a negative correlation because, in general, y decreases as x increases, so that if you were to draw a best-fit straight line through this data, it would have a negative gradient.

1.1.3 No Correlation

And there is said to be *no* correlation between the variables x and y if a set of (x, y) data looks like Figure 3.

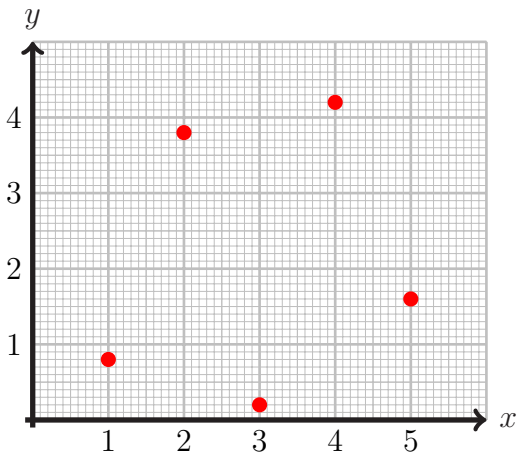

Figure 3: No Correlation

There is no correlation because there is no relationship between the x and y values in this set of data.

1.1.4 The Correlation Coefficient

Mathematicians like to compare things using numbers. Like "this ruler is longer than that one because the length of this ruler is 1 m and the length of that ruler is 30 cm".

Well, so it goes with correlation. In order to compare the correlations of two sets of data, mathematicians assign a number to each set of data. They then just have to compare the numbers to compare the correlations. And the name of the number that they give to a correlation is called the *correlation coefficient*. The correlation coefficient is usually given the symbol r.

There will be more about this later, but here, you just need to know that the value of the correlation coefficient is any number in the range -1 to $+1$. A set of data that has perfect positive correlation will have a correlation coefficient of $+1$; a set of data that has perfect negative correlation will have a correlation coefficient of -1; and a set of data that has no correlation will have a correlation coefficient of 0.

1.2 A Quick Overview of Regression

In the above sections I mentioned an imaginary straight line that we were drawing through our data: the "best-fit" straight line. Well, given a set of data, *regression* is just the process of finding the best-fit line through the data[1].

So, for example, if we were to take the data from Figure 1, then we could draw a best-fit straight line as in Figure 4.

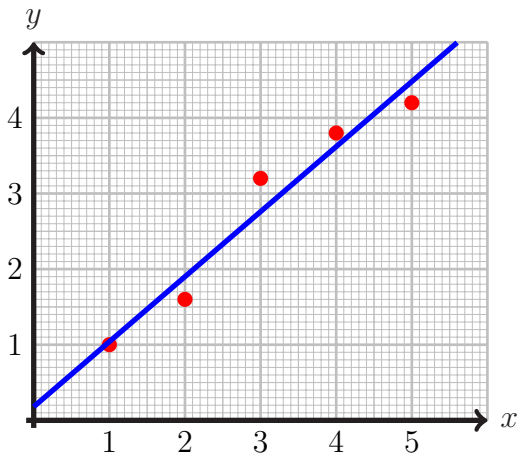

Figure 4: Positive Correlation With Line of Best-Fit

[1]This best-fit line does not have to be straight, it can be any function: a quadratic, cubic, exponential, logarithmic, etc. But in this document we are only worried about straight lines through our data. That's bad enough!

2 Correlation and Regression Equations

2.1 The Product Moment Correlation Coefficient

The bad news: there are as many different correlation coefficients as I've had hot dinners.

The (relatively) good news: there is a standard correlation coefficient that is used by all mathematicians, and it's called the *product moment correlation coefficient* (PMCC).

More bad news: you calculate the PMCC like this:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{1}$$

where

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \tag{2}$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} \tag{3}$$

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \tag{4}$$

and where

$$\sum x = \text{the sum of all the } x \text{ values;} \tag{5}$$

$$\sum y = \text{the sum of all the } y \text{ values;} \tag{6}$$

$$\sum x^2 = \text{the sum of the squares of all the } x \text{ values;} \tag{7}$$

$$\sum y^2 = \text{the sum of the squares of all the } y \text{ values;} \tag{8}$$

$$\sum xy = \text{the sum of: the } x \times y \text{ values for all the data points.} \tag{9}$$

Horrendous, or what?

One sliver of good news is that you don't have to remember any of these equations. They're all in the formula book for your exam. Phew! But you do have to know how to use them.

2.2 The Regression Equations

There's actually plenty of good news here!

Once you have gone through the pain of working out S_{xx}, S_{xy}, and S_{yy} as described in Section 2.1, then actually finding the equation of the best-fit straight line through your data is relatively easy. Here's how to do it.

Remember that the general equation of a straight line is

$$y = mx + c$$

and so it order to determine the equation for our straight line, we need to find the m (the gradient) and the c (the y-intercept).

First, the m:

$$m = \frac{S_{xy}}{S_{xx}} \tag{10}$$

Second, the c. Once we know the gradient, then to find the c, we need a point that the line goes through. And from the theory of all this, it turns out that the best-fit straight line will *always* go through the point

$$(\overline{x}, \overline{y})$$

where

$$\overline{x} = \frac{\sum x}{n} = \text{the average of all the } x \text{ values;} \tag{11}$$

$$\overline{y} = \frac{\sum y}{n} = \text{the average of all the } y \text{ values.} \tag{12}$$

Now remember that the way to find the equation of a straight line if you know the gradient m and a point that the line goes through, (x_1, y_1), is to use the equation

$$y - y_1 = m\,(x - x_1)$$

So the equation of the best-fit straight line will be

$$y - \overline{y} = \frac{S_{xy}}{S_{xx}}\,(x - \overline{x}) \tag{13}$$

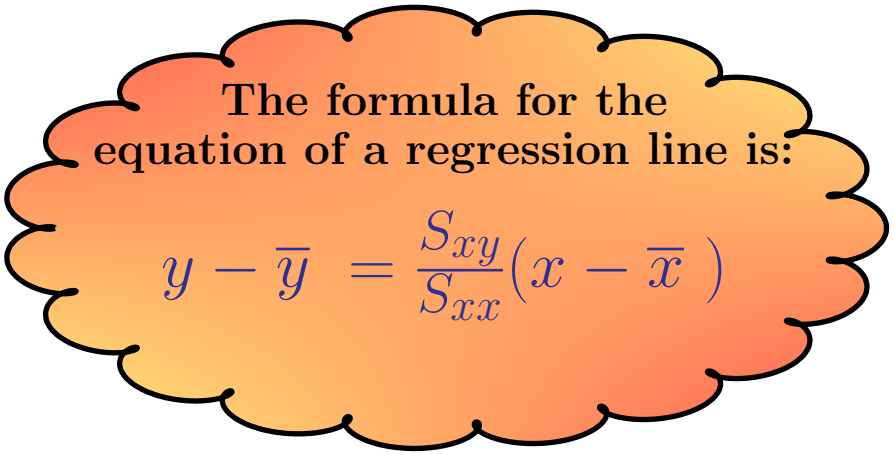

Figure 5: Regression Line Formula

3 Finding Regression Line Equations

Here we're going to find the equations of regression lines in a few different situations. How easy this will be depends entirely on what information you are given...

3.1 You Are Given S_{xy}, S_{xx}, $\overline{x}$ and $\overline{y}$

OK, let's say that in a question you were given this information:

- $S_{xx} = 10$
- $S_{xy} = 8.6$
- $\overline{x} = 3$
- $\overline{y} = 2.76$

Well in that case we can just use the regression formula equation, Equation (13), since we know everything we need from that formula:

$$y - \overline{y} = \frac{S_{xy}}{S_{xx}}(x - \overline{x})$$

so, putting the numbers in:

$$y - 2.76 = \frac{8.6}{10}(x - 3)$$

or, simplifying a bit:

$$y - 2.76 = 0.86(x - 3)$$

Now if we multiply out the brackets:

$$y - 2.76 = 0.86x - 2.58$$

and finally, adding 2.76 to both sides:

$$y = 0.86x + 0.18$$

And there's the equation of the line.

It's very rare to be given this kind of information in an exam question. You are much more likely to get something like...

3.2 You Are Given $\sum x^2$, $\sum xy$, $\sum x$, $\sum y$ and n

This sort of information is very commonly given in exam questions. So if you were given:

- $\sum x^2 = 55$
- $\sum xy = 50$
- $\sum x = 15$
- $\sum y = 13.8$
- $n = 5$

what do you do with that? Well, from the example from Section 3.1 what we need are the values for S_{xx}, S_{xy}, $\overline{x}$ and $\overline{y}$. Well to get those we just use the equations (7), (9), (11) and (12):

$$S_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 55 - \frac{15^2}{5} = 10$$

and

$$S_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 50 - \frac{15 \times 13.8}{5} = 8.6$$

and to get $\overline{x}$ and $\overline{y}$ we use

$$\overline{x} = \frac{\sum x}{n} = \frac{15}{5} = 3$$

and

$$\overline{y} = \frac{\sum y}{n} = \frac{13.8}{5} = 2.76$$

And now we just follow the example of Section 3.1.

3.3 You Are Given the Data Points

What if the only thing that you were given was the data points themselves? Say you were given:

Table 1: Data Points

$\boldsymbol{x}$	1	2	3	4	5
$\boldsymbol{y}$	1	1.6	3.2	3.8	4.2

Now what? Well, from the example of Section 3.2, we can see that the first step in the process is to find $\sum x^2$, $\sum xy$, $\sum x$, $\sum y$ and n.

We can do that directly:

$$n = 5$$

because there are 5 data points. Now,

$$\sum x^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55$$

and

$$\sum xy = (1 \times 1) + (2 \times 1.6) + (3 \times 3.2) + (4 \times 3.8) + (5 \times 4.2) = 50$$

and

$$\sum x = 1 + 2 + 3 + 4 + 5 = 15$$

and

$$\sum y = 1 + 1.6 + 3.2 + 3.8 + 4.2 = 13.8$$

And now we just follow the example of Section 3.2.

3.4 When You Have a Fancy Calculator

I don't want to boast, but I have a CASIO fx-CG 20. It's the best calculator you can get. It does amazing things.

Amongst the many amazing things it does is to perform a wide variety of statistical calculations. For example, if I was given the data from Table 1, I can enter the data into a pair of lists in my calculator by going to

MENU 2

to get me into the Statistical functions of the calculator, then just putting the data into List 1 (the x-values), and List 2 (the y-values). I now press

CALC

to do a calculation on this data, then

REG

to do a *regression* calculation, then

X

then

ax+b

and it gives me (among other things):

$$a = 0.86$$
$$b = 0.18$$
$$r = 0.97425799$$

where the a and the b are the gradient and y-intercept of the regression line respectively, and r is the product moment correlation coefficient.

Job done. Oh, and if you go back to where the Lists are displayed, and press

GRAPH

to draw this data, then

GRAPH1

the calculator will draw the data on a scattergraph. Then press

CALC

then

X

then

ax+b

and it gives me (as before):

$$a = 0.86$$
$$b = 0.18$$
$$r = 0.97425799$$

but then, press

DRAW

and the calculator draws the line of best fit on the scattergraph of the plots.