

r: Regression

The term “regression” refers to one particular way of estimating a summary line to fit a cloud of data. Let me defer the origin of the term “regression” itself: There is a solid reason for the connotation of regressing or moving backward, but the immediate problem is to introduce this particular way of estimating a good line to represent a cloud of data. So far, here’s what you know — and this part does not change: When you have a hypothesis stating that one variable is a linear function of another variable, at least approximately, you generated expected values, the values that your “y” variable would have if the linear hypothesis were correct. Then you examine the exceptions, the residuals, looking at both the size of the residuals and their pattern. You use the size of the residuals both to keep score, how well does the linear hypothesis fit the data, and as a practical device to be used for finding the best slope and the best intercept.

The practical operation for finding the best fit is tedious in the extreme. I’ve used this tedious procedure for two reasons. Primarily, it absolutely forces you to look at the data. And that, in turn, leads to hypothesis construction, to treating a first data point or a last data point as an exception not related to the linear hypothesis, to breaking the curve into pieces, ... to all sorts of intelligent but customized approaches responsive to the problem at hand. Looking at the data, very closely, instead of just committing the data to the computer, programming a set of predetermined questions, and writing up the results, simply leads to better data analysis.

The other reason for the tedious procedure is so that now, when I am prepared to drop the tedium, you will know, nevertheless, exactly what is going on. In most of the examples the procedure was to estimate an intercept and a slope and then compute the mean squared residual

determined by this intercept and slope. Then I re-estimated the intercept and re-estimated the slope and re-examined the mean squared residual, accepting new estimates of the slope where the new estimates reduced this mean squared residual. That procedure is “regression” analysis in everything except name. However it benefits greatly for a straightforward application of the calculus. One thing that the calculus can be very good at is “optimizing”. Instead of the correct but tedious procedure you have been using, we can try to set up a measure of error which is suitable grist for the optimizing devices of the calculus and say, to our calculus, “optimize: find the particular slope and the particular intercept for which the size of the typical residual is reduced to a minimum.” This does not treat the important question of the pattern of the residuals — that remains the proper work for the human eye, and that is why you must look at the graphs even where the calculus can be relied upon to remove the tedium.

The calculus can not optimize just any measure of the residuals but it is excellent at optimizing with respect to the mean squared residual, working in the context of least squares. Here we call upon the calculus to solve a two variable problem, finding a line at the center of a cloud of data. And the procedure is exactly analogous to what we did earlier, finding the center for the values of one variable. Recall that for one variable, the “center” is a point to which the data are close. Using “close” in the sense of least squares implied that the mean was the center of the distribution. Using “close” in the sense of minimum absolute deviation implied that the median was the center of the distribution.

, using close in that is close to the stuff of which it is a center. “Close” can mean different things and the meaning that is historically easiest to work with — when computers are non existent and calculus is well developed — is “close in the sense of least squares”.

More formally, for one variable the average size of the variation around the center (when it is minimized) is called the variance of x :

So the measure of the center that is close to the data in the sense of least squares is the average.

$$V[\$x] = \frac{1}{n} \sum_{i=1}^n (\$x_i - \overline{\$x})^2$$

and the center, defined as the point of minimum variance was the mean.

$$\overline{\$x} = \frac{1}{n} \sum_{i=1}^n \$x_i$$

For interpretive work, when we need a number for spread around the mean that uses the same units as x, we use the standard deviation

$$s_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (\$x_i - \overline{\$x})^2}$$

which is also called by the name “root mean squared error”, specifying the root, the average, and the squares that are visible in the formula.

I am going to use exactly the same logic to find a line which is the best line in the sense of least squares: I will define variation around the line and then choose the line with respect to which that variation is optimized. It is absolutely straightforward. But to make it look absolutely straightforward, I have to return to the use of standardized variables

$$X_i = \frac{x_i - \overline{x}}{s_x} \quad \text{and} \quad Y_i = \frac{y_i - \overline{y}}{s_y}$$

In terms of these standardized forms for x and y it was easy to construct the argument for measuring correlation as

$$r_{xy} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$$

Continuing with these standardized forms of x and y it is straightforward to estimate the best linear summary of the relation between x and y . But remember the trail that leads back from the standardized variable to the original variables: If, for example I learn something about X (upper case “ X ”, the standardized variable), then I have learned something about x (lower case “ x ”, the original variable) with a little algebra to connect one to the other showing that

$$X = \frac{x_i - \bar{x}}{s_x} \quad x = s_x X + \bar{x}$$

which gets us back.

To find the best line, and to let the calculus eliminate the tedium, I construct the linear hypothesis

$$\hat{Y}_i = M X_i + B$$

using a caret, “ \wedge ” over the Y_i to indicate that this is not the true value of Y_i . It is the value that Y_i would have if it were truly predicted from the value of X_i . (with no residual). And now, just as I did for one variable, I create a measure of variation around the center (variation of the Y 's around the line), and prepare to have the calculus find the best line.

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The error variation depends on the choice of the line, which means that error is a function of the intercept B and the slope M . So, I need the

value of B and the value of M for which E is smallest — the best fit line in the sense of least squares.

The details are a combination of algebra and basic calculus. Using the algebra, I want the M and B visible in the equation for error.

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - [M X_i + B] \right)^2$$

And now I use the calculus: The minima of functions occur where their derivatives are equal to zero (or at an end point). So, noting that we are in a peculiar case where M and B are variables (while X and Y are constants -- they are whatever the data make them) — I will differentiate E with respect to B. I differentiate E with respect to M. I will set the two derivatives equal to zero and then solve these two equations as two simultaneous equations in two unknowns, M and B).

— But first, let me establish a few things, “lemmas”, that will make the calculation simple: First I need to establish the average of a standardized variable. If that sounds peculiar to you, if it sounds peculiar to ask for the average of a thing for which I have no data, that’s good. It means you are thinking like a data analyst, which is what I wanted. But one of the mathematically pleasant (and useful) properties of standardized variables is that they have an average, always the same average, as a mathematical fact.

I could tell you the answer, but you should have the habit of proving these things for yourself, there is no need to look them up in the learned text of some expert. How do you get the answer to the question: What is the average of a standardized variable? You simply make the algebraic substitutions and simplify the result. So, starting with the definition of the average, here is the average, for any variable, standardized or not:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Now I make the algebraic substitution, using the definition of the standardized variable.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right)$$

The rest is simplification: I “pull out the common factor” s_x

$$\bar{X} = \frac{1}{s_x} \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \right)$$

and then distribute the summation in order to add up the two terms separately

$$\bar{X} = \frac{1}{s_x} \left(\frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n \bar{x} \right)$$

And now, things are just about reduced: Inside the parentheses, the term on the left is, by definition, the mean of the original variable x .

$$\bar{X} = \frac{1}{s_x} \left(\bar{x} - \frac{1}{n} \sum_{i=1}^n \bar{x} \right)$$

Inside the parentheses, the expression on the right is adding a constant to itself n times, which means it is equal to

$$\bar{X} = \frac{1}{s_x} \left(\bar{x} - \frac{1}{n} n\bar{x} \right)$$

which simplifies to

$$\bar{X} = \frac{1}{n} (\bar{X} - \bar{X})$$

which simplifies further to the answer

$$\bar{X} = 0$$

There is the result I need now in order to make subsequent computations simple. It tells me that any time I see \bar{X} in an equation I can substitute the value 0.

That is peculiar if you are thinking about data: Here is a variable whose mean is always zero. But, of course, it is a standardized variable that was created expressly for the purpose of having a variable that was centered on its own average. All I've done with the mathematics is recover this fact about standardized variables by "proving" that the average (of a standardized variable) is 0.

In the same spirit, what is the variance of a standardized variable? If there is any doubt about the answer, I figure it out, as before, by making substitutions and simplifying the result. I begin with the definition. The variance of any variable, standardized or not, is

$$S_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

Using what I have just established about the mean of X, I use what I established to simplify the present equation.

$$S_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i)^2$$

Now I am ready for substitution, using the definition of X.

$$S_X^2 = \frac{1}{n} \sum_{i=1}^n \frac{X_i - \bar{X}}{S_X}^2$$

This invites simplification by factoring-out the denominator within the parentheses

$$S_X^2 = \frac{1}{S_x^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

And now if you will check your definitions, you will recognize the thing within the square brackets on the right. By definition, it is the variance. So:

$$S_X^2 = \frac{1}{S_x^2} [S_x^2]$$

From which it follows

$$S_X^2 = 1$$

There is the second result I need now in order to make subsequent computations simple. It tells me that any time I see S_X^2 in an equation I can substitute the value 1. (

With these two lemmas in hand, I am ready to work on the partial derivatives of $E(M,B)$, to find the best fit line in the sense of least squared error with respect to the variable Y .

I will differentiate E with respect to B and E with respect to M , producing two expressions. Eventually I will set both expressions equal to 0 creating two simultaneous equations that I will solve for B and M . But first, I simplify. Substituting the formula for E , I write

$$\frac{\partial}{\partial B} E(M, B) = \frac{1}{B} \sum_{i=1}^n (Y_i - [MX_i + B])^2$$

$$\frac{\partial}{\partial M} E(M, B) = \frac{1}{M} \sum_{i=1}^n (Y_i - [MX_i + B])^2$$

Knowing that the derivative of the sum is the sum of the derivatives, the pair of expressions becomes

$$\frac{\partial}{\partial B} E(M, B) = \frac{1}{B} \sum_{i=1}^n \frac{\partial}{\partial B} (Y_i - [MX_i + B])^2$$

$$\frac{\partial}{\partial M} E(M, B) = \frac{1}{M} \sum_{i=1}^n \frac{\partial}{\partial M} (Y_i - [MX_i + B])^2$$

Using the chain rule to deal with the squares,

$$\frac{\partial}{\partial B} E(M, B) = \frac{1}{B} \sum_{i=1}^n 2(Y_i - [MX_i + B]) \frac{\partial}{\partial B} (Y_i - [MX_i + B])$$

$$\frac{\partial}{\partial M} E(M, B) = \frac{1}{M} \sum_{i=1}^n 2(Y_i - [MX_i + B]) \frac{\partial}{\partial M} (Y_i - [MX_i + B])$$

Only one of the expressions inside the nested parentheses at the right has the variable B in it, and only one has an M, so the derivatives simple again, leaving

$$\frac{\partial}{\partial B} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n (Y_i - [MX_i + B])$$

$$\frac{\partial}{\partial M} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n (Y_i - [MX_i + B]) X_i$$

Eventually, I have to set both of these expressions equal to zero and solve the simultaneous equations for B and M. But first, I use the lemmas to simplify these expressions.

Recalling that the derivative of the sum (within the parentheses) is the sum of the derivatives And noting that differentiating with respect to B is particularly simple I get

$$\frac{\partial}{\partial B} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n Y_i - \frac{1}{n} \sum_{i=1}^n MX_i - \frac{1}{n} \sum_{i=1}^n B$$

$$\frac{\partial}{\partial M} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n MX_i^2 - \frac{1}{n} \sum_{i=1}^n BX_i$$

factoring out some the M's and the B's

$$\frac{\partial}{\partial B} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n Y_i - M \frac{1}{n} \sum_{i=1}^n X_i - B \frac{1}{n} \sum_{i=1}^n 1$$

$$\frac{\partial}{\partial M} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n X_i Y_i - M \frac{1}{n} \sum_{i=1}^n X_i^2 - B \frac{1}{n} \sum_{i=1}^n X_i$$

That shows recognizable terms which are means and variances of standardized variables

$$\frac{\partial}{\partial B} E(M, B) = -2 \left(\bar{Y} \right) - M \left(\bar{X} \right) - B \frac{1}{n}$$

$$\frac{\partial}{\partial M} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n X_i Y_i - M \left(s_X^2 \right) - B \left(\bar{X} \right)$$

That allows a sharp simplification to

$$\frac{\partial}{\partial B} E(M, B) = 2B$$

$$\frac{\partial}{\partial M} E(M, B) = -2 \frac{1}{n} \sum_{i=1}^n X_i Y_i - M$$

And the one complicated looking term is also recognizable. The context is different, but this is the “r” that has been used, earlier, as a measure of correlation.

$$\frac{\partial}{\partial B} E(M, B) = 2B$$

$$\frac{\partial}{\partial M} E(M, B) = -2(r - M)$$

Now I go back to the game plan: I set these two partial derivatives equal to zero.

$$2B = 0$$

$$-2(r - M) = 0$$

Now, finally, I am rarely going to find a pair of simultaneous equations more easily solved than these. The first tells me $B=0$. The second tells me $M=r$.

$$\begin{cases} B = 0 \\ M = r \end{cases}$$

(** Figure out how to get the Equation editor to align those equations to the left.)

Don't miss the simplicity of these statements: They say that B is always 0. The intercept of the best line for standardized X and standardized Y is *always* 0 — as you saw it in the graph where the line passed at least close to zero.

And while the slope is not simple, it shouldn't be: Something in this equation ought to depend on data. It is the slope. And while it isn't simple it is very nice to see "r" recurring. Previously r was introduced as a measure of correlation. Here, the same r is the slope of the best fit line in the sense of least squares for predicting Y as a linear function of X:

The fruit of all the standardization and all the math is that, in this form, it is simple. In standardized form the best line (best in the sense of least squares) is always

$$Y = rX$$

And now, back to data: What does this statement about X and Y say about the data variables, x and y, with which I began this odyssey into the optimization of a straight line. The formula for the best line for x and y is more complicated, but I don't have to remember it. I just remember the standardized equation and make the substitutions. \

Starting with the basic equation:

$$\widehat{Y}_i = r X_i$$

and undoing the standardization

$$\frac{\hat{y}_i - \bar{y}}{s_y} = r \frac{x_i - \bar{x}}{s_x}$$

Or

$$\left(\hat{y}_i - \bar{y}\right) = r \frac{s_y}{s_x} \left(x_i - \bar{x}\right)$$

Cleaning it up, to show x multiplied by a slope and to show what is added (the intercept) at the end (enclosing the slope and the intercept in parentheses

$$\hat{y}_i = \left(r \frac{s_y}{s_x} \right) x_i + \left(\bar{y} - r \frac{s_y}{s_x} \bar{x} \right)$$

That gives you the best slope and the best intercept, best in the sense of least squares. Take your choice, tedious minimization on the spread sheet or one shot, here's the answer, using the calculus to tell you how to convert the means, standard deviations, and the correlation coefficient r into the best answer.

Variance of Data = Variance of Signal + Variance of Noise:

How good is the best?

We're not quite done: When I asked for the center of the data, center in the sense of least squares, I asked two questions. First I asked which center was the best. Then I asked, "How good?" and got the variance and the standard deviation as the answers. Here, first I asked what was the best line, best in the sense of least squares. Now I ask "How good?" And I will get the variance and standard deviation of the residuals (whose average is zero). In addition, there is a fringe benefit: The answer to that question is so "nice" in a mathematical sense that it leads to certain conventional presentations. It is not obvious that the mathematical "niceness" of the least squares method should weigh heavily among the priorities of the data analyst choosing a method but, in use, it does.

So, resuming the discussion. In order to choose a best fit line, best in the sense of least squares, here is what I minimized:

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

How big is it, at its minimum? To answer the question, I just substitute what I know

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n (Y_i - [rX_i])^2$$

and I simplify it — continuing to known attributes of standard variables (mean 0, variance 1).

Squaring

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n \left(Y_i^2 - 2rX_i Y_i + (rX_i)^2 \right)$$

Distributing the summation through the terms

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n Y_i^2 - 2r \frac{1}{n} \sum_{i=1}^n X_i Y_i + r^2 \frac{1}{n} \sum_{i=1}^n X_i^2$$

And in this form I see the now-familiar terms. The term on the left uses the variance of a standardized variable. That's one. The term on the right uses the variance of a standardized variable. That's one. And the term in the center uses the mean product, for which we have the symbol r . So

$$E(M, B) = 1 - 2rr + r^2$$

And that is the size of the error: It is always

$$E(M, B) = 1 - r^2$$

Once again, the story is told by r . r is the measure of correlation. r is the slope of the regression line in standard form. And, now, r is the key quantity in assessing the size of the error.

The fringe benefits of this equation lead to mathematical nice results. For example, from this equation for the size of the error, it follows that r has absolute limits of minus one and plus one. This follows because I know one thing for sure about the *squared* error: Squared error can not be negative. Therefore zero is less than or equal to the error

$$0 \leq 1 - r^2$$

And that equation tells me about the limits of correlation: Because zero is less than or equal to one minus r-squared,

$$r^2 \leq 1$$

And as a consequence r itself is bounded by the interval plus-or-minus 1

$$-1 \leq r \leq 1$$

I can also figure out the limits of a bad correlation. How weak can a correlation be? You would like the English, “no correlation” to correspond to a mathematics that says “zero correlation”. But if it is true, it has to be proved. So, suppose that X doesn’t help at all as a predictor of Y? Suppose that I always predict \bar{Y} regardless of X. Suppose that whatever the value of X, I always predict that Y will be equal to its mean — the number that is close to all the y’s, but gets no help from x? In this sad case the error is

$$E(M, B) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

This formula is recognizable: Assuming that the prediction is always \bar{Y} , this formula for the error is identical to the variance of the standardized variable Y. The error is σ^2 . And since the error is 1, I substitute this value into the equation and it determines a value for r. Thus, in the case that X is useless for predicting Y,

$$1 = 1 - r^2$$

And

$$r = 0$$

That gives me the value of r for no correlation. I can also ask for the value of r when the line is a perfect fit to the data (no non-zero residuals). The value of r had better be plus or minus 1, but again I have to prove it. So, suppose that the error vanishes

$$0 = 1 - r^2$$

Sure enough, solving that equation for r , r must be plus or minus 1.

And finally, one more piece of the puzzle that leads to the standard jargon: Finally, I want to look at the three pieces: Data, Signal, and Noise and show that the pieces add up. I already know

$$\text{Data} = \text{Signal} + \text{Noise}$$

That was an assumption. Now I am going to demonstrate something that is not an assumption. I am going to demonstrate that the *variance* of the data is equal to the *variance* of the signal plus the *variance* of the noise. When I can demonstrate that I will be able to phrase sentences that sound very good to the data analyst. I will be able to say how much of the data is explained by the hypothesis about the signal — which means “How much of the variance of the data is the variance of the hypothetical signal?” or letting the language drift into conventional form “What percent of the variance is *explained* by the hypothesis?”

The first step is to demonstrate the equation

$$\text{Variance of the Data} = \text{Variance of the Signal} + \text{Variance of the Noise}$$

(where “variance of the signal” means variance of the values that would be predicted using a linear equation for y as a function of x).

The variance of the data is simply the variance of Y, the variable we are trying to describe as a function of X. So

$$\text{Variance of Data} = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

Now let me add and subtract a useful term that leaves the sum unchanged:

$$\text{Variance of Data} = \frac{1}{n} \sum_{i=1}^n (Y_i - rX_i + rX_i)^2$$

Regrouping

$$\text{Variance of Data} = \frac{1}{n} \sum_{i=1}^n ((Y_i - rX_i) + rX_i)^2$$

and then squaring, I get three pieces

$$\text{Variance of Data} = \frac{1}{n} \sum_{i=1}^n (Y_i - rX_i)^2 + 2r \frac{1}{n} \sum_{i=1}^n X_i (Y_i - rX_i) + \frac{1}{n} \sum_{i=1}^n (rX_i)^2$$

Now two of these three pieces look familiar: The term on the left is the variance of the residuals. The term on the right, with components rX_i , is the variance of the predicted values of Y (around an average prediction of zero). That leaves "stuff" in the middle which, we hope, is zero.

Variance of Data = Variance of Error + stuff + Variance of Signal

But it must be checked. So, starting at $1/n$

$$\frac{1}{n} \sum_{i=1}^n X_i (Y_i - rX_i) = \frac{1}{n} \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n rX_i^2$$

I recognize both of these expressions, factoring r from the second term

$$\frac{1}{n} \sum_{i=1}^n X_i (Y_i - rX_i) = r - r \frac{1}{n} \sum_{i=1}^n X_i^2$$

from which it follows

$$\frac{1}{n} \sum_{i=1}^n X_i (Y_i - rX_i) = r - r$$

and

$$\frac{1}{n} \sum_{i=1}^n X_i (Y_i - rX_i) = 0$$

This gets rid of the middle term, its value is zero, and leaves the target equation

$$\text{Variance of Data} = \text{Variance of Error} + \text{Variance of Signal}$$

or

$$1 = (1-r^2) + r^2$$

where

the Variance of Data is 1

the Variance of Error is $(1-r^2)$

and

the Variance of Signal is r^2

or in another technical expression for the same thing

$$\text{Total Variance} = \text{Unexplained Variance} + \text{Explained Variance}^1$$

This leads to phrases that you will read over and over again in statistical reports, phrases like “The correlation explains 40% of the variance.” The phrase refers to the fact that the total variance is 1 and that the unexplained variance and the explained variance add up to one. so if the variance of the predicted values is .40 and the variance of the residuals is .60, you may say (with somewhat dubious linguistic precision), 60 percent of the variation is unexplained. Or, the prediction explains 40 percent of the variance.

¹ Over and over again here I am on thin ice referring to these things as variances: After all, variances are variations around a mean, and I am not showing a mean in this formula. To clean it up, I have to show that the missing mean is zero.

Exercises

1. Look at the data for Brain Weight and Body Weight. Write a short statement describing the quality of the prediction in conventional terms, using r and r^2 .

Now explore the limits of the conventional claims that one variable “explains” another:

2. Repeat the first problem using brain weight and body weight, without logs. Compare this to the first answer (using logs). The answers are different. Reconcile them.

3. Compute some hypothetical data where your x goes from 1 to 10 and your y increases multiplicatively (below). You know for certain that these y 's are not a linear function of these x 's — in this case $y = 1.1^{(x-1)}$. But, as a thought experiment, be dumb: Use a linear equation, use regression, to predict y from x . Use r and r^2 to report how well x explains y . Reconcile the superficial implications of the r and r^2 with the fact that anyone claiming that this y is a linear function of this x has clearly failed to explain the relation at all. (The word “explain” is rich with ambiguity. It has many meanings.)

X	Y
1	1
2	1.1
3	1.21
4	1.331
5	1.4641
6	1.61051
7	1.771561
8	1.9487171
9	2.14358881
10	2.35794769

--