Fitting causal models to data on the number of papers in each field using linear regression

Sarah Marzen

February 6, 2025

Abstract

The point of this lab is to fit causal models to data, where the constants in the models describe how papers spawn new papers and mutate into another field. By the end of the lab, we will ask: are the models discussed in class [1] good models for describing our data, generously scraped by my husband from websites with repositories of papers, on how many papers there are in each field as a function of time?

1 Instructions for turning in your labwork

I will ask that you turn in a Word document that answers all bolded questions, potentially with plots in tow if requested. (For that, screenshots are fine, but copy and pastes are preferred.) All answers are fine– they just have to be supported by the data.

2 Fitting linear equations to data

First, we will prove to ourselves that we can recover the best-fit linear model from data– or examine when that does not quite work. In single-variable linear model, the dependent variable x is related to the independent variable y via the equation

$$y = mx + b + \eta, \tag{1}$$

where η is "noise" that corrupts our observations or additional factors that we haven't measured. We would like to show that, from data, we can figure out m and b– the slope and the intercept– or in other words, the model that describes the data.

2.1 Build your own simulated data

The first step in any scientific endeavor is checking that your method works. The entire method to understand the more complicated data on number of papers as a function of time [1] is based on linear regression. We first want to show that our method (linear regression) works on data that we understand "ground truth"– the answer– for. We'll then apply it to the new data, confident it works in at least one test case.

If you've never seen it before, in linear regression, the idea is that the dependent variable depends is some constant multiplied by the independent variable added to a constant plus some noise in single-variable linear regression. If you want, you can have multiple independent variables, all multiplied by constants, all added together plus a constant plus some noise in multivariate linear regression, which we will do later. But, just to get a feel, we'll start with single-variable linear regression.

In Excel, label one sheet as "Simulated data". Label one column, column A, as "IV" for "independent variable". So, A1 should say "IV". Label the next column, column B, as "DV" for "dependent variable". So B1 should say "DV".

Now we make up our data.

To make it easy and still rigorous enough, let's just populate column A with the integer that corresponds to its row. To do so, set A2 to 2. Then set A3 to be the formula "=A2+1", where $A2$ means that you actually click on A2. Hit enter. You should see that A3 is now 3. Then, click and drag A3 all the way down to A30. You should see that the column A is now populated with the row number all the way down to a count of 30 at A30.

Now, let's make up some simulated data for IV in column B that is linear. For cell B2, set it to "$= A2+1+NORMINV(RAND(), 0, 1)$". This will make $m = 1$, $b = 1$, and η be a certain kind of noise that linear regression is well-adapted to handle. Here, "NORMINV(RAND(),0,1)" is the noise. This should give you a value for B2 that may be different than your neighbor. That is okay. Click and drag B2 all the way down to B30. The numbers should increase, but may not always increase– sometimes there may be a decrease in column B from one cell to the next. To keep the numbers from changing, select all the cells in the IV column except for B1 and hit "Copy". Then click "Paste Special" and "Values".

This is your random data! Plot it to see what it looks like. To do this, grab column A and column B and click on the "Insert" tab. You should see an option for plotting with one plot that looks like a bunch of points randomly scattered. With column A and column B still highlighted, click on the "scatter" option. You should see a bunch of points that look like they follow a line with some noise. This is the simulated data! **Put this plot in a Word document and explain what's going on. What is the x-axis and what is the y-axis?**

2.2 How to fit to a line, roughly speaking

We will now start making guesses for what the true model of the data is, pretending that we do not know the answer. This is a crucial part of computational science. You first generate numbers where you know the answer and see if your method can recover the answer that you know to be true, at least approximately.

In column C, write in C1 "Guess". In C2, write "$= 0.5 * A2 + 0.25$" as our starting guess. Click and drag C2 all the way down to C30. Numbers should populate that keep on increasing, but may not match column B.

In other words, we are *arbitrarily* starting our guess of what is correct by guessing that $m = 0.5$ and $b = 0.25$. Let's see how we can figure out if this is remotely right.

First, let's do it using a picture. Delete your scatter plot from the previous subsection and remake it with column A as the x-axis, and *both* column B and column C on the y-axis. To do this, grab all three columns and repeat the process to create a scatter plot. You should see a scatter plot with blue dots for the real simulated data and orange dots for our guess as to the value of the dependent variable based on the independent variable and our current model. You should see from the plot that the orange dots are off the blue dots by quite a bit!

But how can we see this numerically? In column D, say that D1 is "MSE", which stands for "mean-squared error". In D2, write "$= (C2 - B2) \wedge 2$" and click Enter. You should see that D2 is a positive number that shows how far away C2 is from B2. The farther away the two values are, the farther away that guess is from the actual independent variable, the larger D2 will be. Click and drag D2 all the way down to D30. You should see a list of numbers that likely grows, potentially to be quite large. In D31, write "$= AVERAGE(D2 : D30)$". This is a number that is large if all of the guesses are far away, medium if some of the guesses are far away, and small if the guesses are close to the true simulated data. The meaning of "large", "medium", and "small" depends on how close we really can get to the simulated data, but we will get into that in just a second. **In the Word document, show your new plot and write the value of the mean-squared error beneath it. Explain as best you can what it means.**

Now for the part that explains what modelers do– keep on changing the m and b in the "Guess" column until the mean-squared error is as low as you can make it. To do so, go back to the formula in C2 and change it whatever you want. So if you want an m of 0.75 and a b of 0.1 and want to try that out, change the formula to "$= 0.75 * A2 + 0.1$ and click and drag until you get all the way down to C30. Try at least five different m's and b's, trying to minimize mean-squared error as much as possible. **In your Word document, for each combination of m and b, report the mean-squared error and show the corresponding plot of the true DV and the guesses. Hint: m should be close to 1 and b should be close to 1. Why?**

2.3 Getting Excel to do the work for you

Now that we've gone through a painstaking procedure to try and find the right m and b– the right model for our data– we will ask Excel to do it for us.

In column E, in cell E1, write "Slope". In cell E2, write "$= SLOPE(B2 : B30, A2 : A30)$". A number should appear that is close to 1. This is Excel's best guess at m. In cell E3, write "Intercept". In cell E4, write "$= INTERCEPT(B2 : B30, A2 : A30)$". A number should appear that is close to 1. This is Excel's best guess at b.

Excel is doing exactly what you were doing, but it is searching through all m and b automatically using a computer program. There are actually closed-form expressions if you know multivariable calculus for the answer to what m and b should be guessed as based on this idea of searching through all possible m and b, and Excel does that for you.

In column F, in cell F1, write "Excel's guess". Then populate column F with Excel's predictions for the independent variable. In other words, using your slope m and intercept b from Excel in column E, write that cell F2 is $mx + b$ and click and drag down to cell F30. **Make a new scatter plot that shows the true dependent variable and Excel's guess versus the independent variable and report Excel's m and b. Use column F to also report Excel's mean-squared error. Is it lower than yours?**

Finally, if we want to measure how linear our data is, there's a way to do so. Mean-squared error does tell us how far away we are from the line, but a *squared correlation coefficient* normalizes that by the spread of the data itself and subtracts from 1. *This is probably the most important number you should retain from doing this lab.* The closer this value is to 1, the closer the data is to a perfect noiseless line– the more correlated it is. This line could be such that increases in the dependent variable lead to decreases in the independent variable or that increases in the dependent variable lead to increases in the independent variable; the only question is how *reliably* changes in the dependent variable lead to predictable changes in the independent variable. The more reliable the changes, the higher the squared correlation coefficient. To find this value using Excel, type "R Squared" into cell E5 and type "$= CORREL(A2 : A30, B2 : B30) \wedge 2$" into cell E6. **Report this value in your lab write-up. How close is it to 1, and what does that mean?**

3 Building a linear model of real data on how many papers are in each scientific field

My husband, a data engineer, scraped a repository (OpenAlex iSearch) for how many books, theses, and papers in a given field were published in the repository every year. I'm giving you, in an Excel file on Canvas, the data for several fields (physics, chemistry, biology, psychology, math, engineering, computer science, and environmental analysis) from the year 1970 until 2024. Grab this Excel file, and let's get going! We're going to be fitting a simple model to this data to describe the number of papers produced every year.

3.1 Reshaping historical data for fitting models

We will now make a model of the number of physics papers in the year $t+1$ based on the papers published so far in each field by year t. Let's label our variables:

- ph_t is the number of physics papers published in year t, and Ph_t is the number of physics papers published total by the end of year t;
- c_t is the number of chemistry papers published in year t, and C_t is the number of chemistry papers published total by the end of year t;
- b_t is the number of chemistry papers published in year t, and B_t is the number of biology papers published total by the end of year t;
- e_t is the number of engineering papers published in year t, and E_t is the number of engineering papers published total by the end of year t;
- m_t is the number of math papers published in year t, and M_t is the number of math papers published total by the end of year t;

- cs_t is the number of computer science papers published in year t, and Cs_t is the number of computer science papers published total by the end of year t;
- p_t is the number of psychology papers published in year t, and P_t is the number of psychology papers published total by the end of year t;
- ea_t is the number of environmental analysis papers published in year t, and Ea_t is the number of environmental analysis papers published total by the end of year t.

Note that Ph_t, C_t, B_t, E_t, M_t, Cs_t, P_t, Ea_t are all the columns between K and R, while ph_t, c_t, b_t, e_t, m_t, cs_t, p_t, ea_t are all the columns between B and I.

Please plot the year versus the number of papers in each field, with each field a different color, in Excel, and add this file to your lab report. Use the scatterplot function described earlier. *If* you also grab the first row that has labels and no numerical data, the graph should be automatically titled on the x and y axes with the legend filled out. What do you notice about this hard-scraped data? Do you trust it? Do these numbers make sense, or is this repository maybe not as complete as it could be? What's your guess, and how would you sanity check?

3.2 Fitting the basic creativity model

Our model for the number of new physics papers every year will be that:

$$ph_t = w_{Ph}Ph_t + w_C C_t + w_B B_t + w_E E_t + w_M M_t + w_{Cs} Cs_t + w_P P_t + w_{Ea} Ea_t + b + \eta_t, \qquad (2)$$

where η_t is noise that is due to many other factors we are not modeling. The coefficients $w_{Ph}, w_C, w_E, w_M, w_{Cs}, w_P, w_{Ea}$ and intercept b (the additional papers that are published every year regardless of how many papers are already in existence) are what we want to find. And additionally, we would like to figure out how well this simple model captures the dynamics of how much scientific fields explode. The idea behind this model is that papers that already exist in other fields spark the idea for new papers in physics. Is that at all correct?

(At this point, you may ask: why concentrate on physics? Well, this is a physics class!)

To figure out what these coefficients are, what the intercept is, and how well the model works, we are going to need to activate a Data Analysis Toolkit in Excel. Please click on the Data tab and then click on Analysis Tools. Check that you want the "Analysis ToolPak" add-in. Call me over if this does not work.

To fit this multivariate linear model, which works on the same idea that you've already played with, we can just get Excel to do the legwork for us. In the Data tab, click on "Data Analysis". (If that button isn't there, try making sure you have the "Analysis ToolPak" add-in checked, and call me over with any problems.) Click on "Regression", and hit "OK". Your "Input Y Range" should be the new physics papers, starting from year 1971, all the way to 2024– "B2 : B56". For your "Input X Range", grab "K2 : R56". Check "Labels", do *not* check "Constant is Zero" because we are allowing for a nonzero b, and click the button next to "Output Range". Set it to T2. Then click ok and watch the data come back.

You should have a beautiful table laid out for you with coefficients and intercepts. The "intercept" is b. Its standard error comes from resampling the data and asking if the intercept changes much if you choose only a subset of the data. The coefficients correspond to the columns that you chose as independent variables, in order. There should be labels in the table that Excel makes for each of the coefficients that will allow you to figure out which number is which.

Report the coefficients $w_{Ph}, w_C, w_B, w_E, w_M, w_{Cs}, w_P, w_{Ea}$ and intercept b and the "R Square" value, which is the earlier-described squared correlation coefficient. How close is it to 1?

3.3 Conclusions

Let's try to figure out if this simple model fit the data well.

Comment on the closeness of the squared correlation coefficient to 1 and assess if this model is likely, therefore, to be a good model of the data. Were you surprised by how well the model did, or did you think it could have been way better? If so, how could we do a better job? Answer in only a paragraph.

It turns out that these multivariate regressions are likely to return incorrect values for parameters like the coefficients $w_{Ph}, w_C, w_E, w_M, w_{Cs}, w_P, w_{Ea}$ and intercept b if the so-called independent variables are actually highly correlated. The model is then still predictive, but it does not actually get the right values for the coefficients and intercept. Speculate on whether or not you think this is a problem for us. **Do you think those coefficients and intercept reflect reality, or is this just an expression that can fit the data just because sometimes linear combinations of things fit data randomly well? For example, are we predicting that increases in papers in psychology lead to decreases in the number of physics papers in the next year– would that make any physical sense?**

3.4 Extra credit

Next, let's make one final graph for extra credit, if you want. **If you want to, for extra credit, make a new column in Excel which is the predicted number of physics papers in every year from the linear model that we fit. Use the coefficients and intercept in the table to populate this column, and scatter plot both the true number of papers every year in physics and the predicted number of papers every year in physics versus the year. Paste this graph into your lab report. Based on this final graph, would you say that the model fit the data well?**

References

[1] Baciu, Dan C. "Causal models, creativity, and diversity." *Humanities and Social Sciences Communications* 10.1 (2023): 1-15.