Does simple linear regression imply causation?

Asked 11 years, 4 months ago Modified 4 months ago Viewed 60k times

21

I know correlation does not imply causation but instead the strength and direction of the relationship. Does simple linear regression imply causation? Or is an inferential (t-test, etc.) statistical test required for that?

regression correlation causality

5

Share Cite Edit Follow Flag

edited May 12, 2011 at 6:44
chl
51.5k 18 207 369

asked May 11, 2011 at 23:05
user4572
211 1 2 3

3 What do you mean by "direction"? Have you read the answers to similar questions stats.stackexchange.com/search?q=causal ? The short answer is no! – NRH May 11, 2011 at 23:41

3 Neither of your suggestions imply causation (or direction). – Henry May 11, 2011 at 23:43

2 I think the OP meant "direction" in the sense of positive vs negative correlation, not the direction of any causal relationship between X and Y. – JMS May 13, 2011 at 21:39

I read all the answers below. Some insight are useful but no one answer seems me decisive. I offered an answer about “Regression and causality” here (stats.stackexchange.com/questions/493211/... think that it give decisive answer to the question. – markowitz Nov 16, 2020 at 11:51

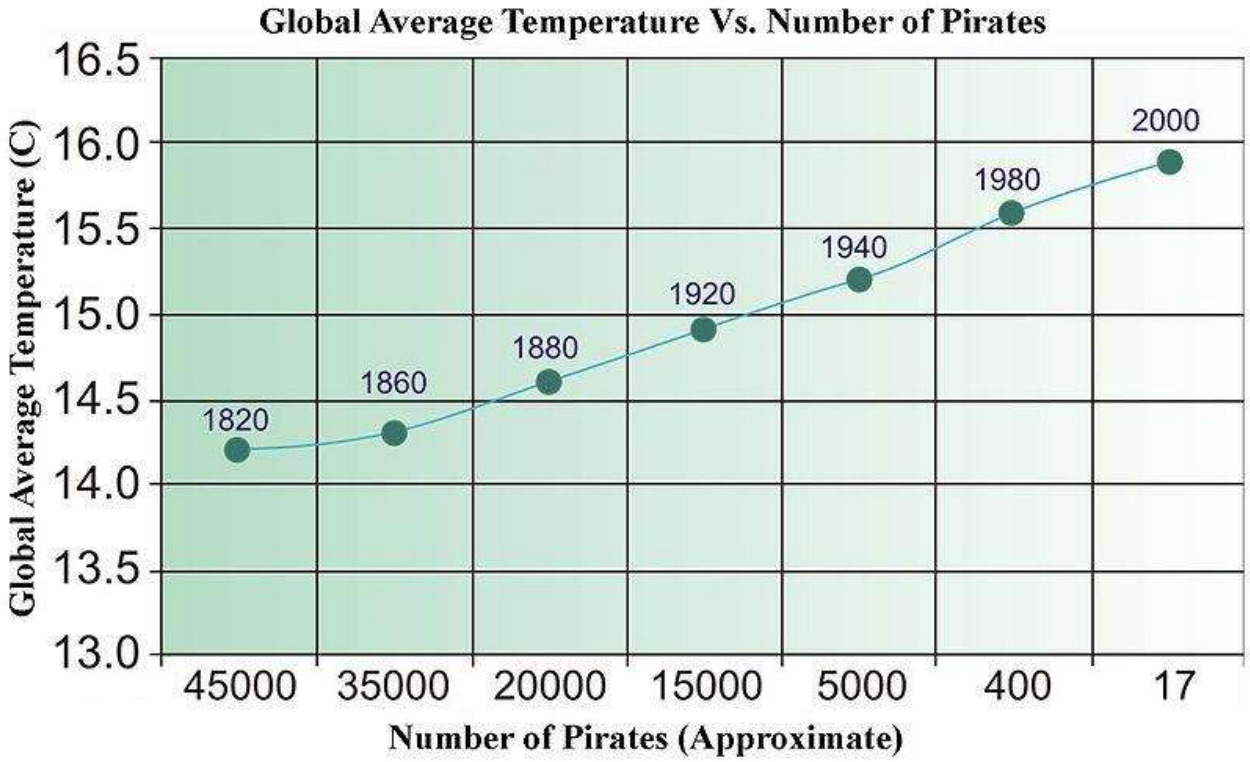
6 Answers

Sorted by: Highest score (default)

23

The quick answer is, no. You can easily come up with non-related data that when regressed, will pass all sorts of statistical tests. Below is an old picture from Wikipedia (which, for some reason has recently been removed) that has been used to illustrate data-driven "causality".

We need more pirates to cool the planet?



For time series, there is a term called "Granger Causality" that has a very specific meaning.

http://en.wikipedia.org/wiki/Granger_causality

Other than that, "causality" is in the eye of the beholder.

Share Cite Edit Follow Flag

edited May 11, 2011 at 23:57

answered May 11, 2011 at 23:44

bill_080
3,478 1 20 21

I meant positive correlation or negative by direction. Thankyou for your response and link to similar questions. – user4572 May 12, 2011 at 0:49

2 Thats quite a crazy X axis in that picture! (But good example!) – Andy W May 12, 2011 at 1:47

2 Another.....Cheese, Butter, and Sheep in Bangladesh, versus the S&P500 (R^2=0.99) ..... nerdsonwallstreet.typepad.com/my_weblog/files/... .... – bill_080 May 12, 2011 at 2:02

5 That graph is *obviously* outdated. Either that or there is bias due to lack of surveyors available to sample in the Gulf of Aden – cardinal May 12, 2011 at 13:23

2 That data was before Al Gore became a pirate. – bill_080 May 12, 2011 at 15:01

12 There is nothing explicit in the mathematics of regression that state causal relationships, and hence one need not explicitly interpret the slope (strength and direction) nor the p-values (i.e. the probability a relation as strong as or stronger would have been observed if the

the probability a relation as strong as or stronger would have been observed if the relationship were zero in the population) in a causal manner.

That being said, I would say regression does have a much stronger connotation that one is estimating an explicit directional relationship than does estimating the correlation between two variables. Assuming by correlation you mean Pearson's r, it typically does not have an explicit causal interpretation as the metric is symmetrical (i.e. you can switch which variable is X and which is Y and you will still have the same measure). Also the colloquialism "Correlation does not imply causation" I would suspect is so well known that stating two variables are correlated the assumption is one is not making a causal statement.

Estimated effects in regression analysis are not symetrical though, and so by choosing what variable is on the right hand side versus the left hand side one is making an implicit statement unlike that of the correlation. I suspect one intends to make some causal statement in the vast majority of circumstances in which regression is used (inference vs prediction aside). Even in cases of simply stating correlations I suspect people frequently have some implied goals of causal inference in mind. Given some constraints are met correlation can imply causation!

Share Cite Edit Follow Flag

edited Apr 13, 2017 at 12:44
Community Bot
1

answered May 12, 2011 at 5:15
Andy W
15.4k 8 74 192

7

Neither correlation nor regression can indicate causation (as is illustrated by @bill_080's answer) but as @Andy W indicates regression is often based on an explicitly fixed (i.e., independent) variable and an explicit (i.e., random) dependent variable. These designations are not appropriate in correlation analysis.

To quote Sokal and Rohlf, 1969, p. 496

> "In regression we intend to describe the dependence of a variable *Y* on an independent variable *X*... to lend support to hypotheses regarding the possible causation of changes in *Y* by changes in *X*..."
>
> "In correlation, by contrast, we are concerned largely whether two variables are interdependent or *covary* - that is, vary together. We do not express one as a function of the other."

Sokal, R. R. and F. J. Rohlf, 1969. *Biometry.* Freeman and Co.

Share Cite Edit Follow Flag

edited Jun 11, 2020 at 14:32
Community Bot
1

answered May 12, 2011 at 11:56
DQdIM
1,079 2 9 20

From a semantic perspective, an alternative goal is to build evidence for a good predictive model instead of proving causation. A simple procedure for building evidence for the

model instead of proving causation. A simple procedure for building evidence for the predictive value of a regression model is to divide your data in 2 parts and fit your regression with one part of the data and with the other part of the data test how well it predicts.

4

The notion of Granger causality is interesting.

Share Cite Edit Follow Flag

answered May 12, 2011 at 18:16

b_dev
931 5 11

Prediction does not imply causality. Best predictive models are autoregressive ones. Zero explanatory power. – luchonacho Sep 9, 2021 at 19:44

If you think of regression coefficients:

$$\hat{\beta} = \frac{Cov(x,y)}{Var(x)}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

3

where Var(.) and Cov(.) are estimates from sample (data).

Consequently, these parameters themselves are nothing else than some functions of correlation between x and y. Especially, beta is just a "normalised" correlation coefficient. So, theres is no more implied causality in regression than in correlation. Causal regression is a special technique in econometrics where one would have to rely on e.g. instrumental variables to get around phenomenons like confounding that obscure the causal interpretation of any particular regression model.

My point is: regression can be *made* causal but it is *not* causal y default.

For more see these videos: https://www.youtube.com/watch?v=Sqy_b5OSiXw&list=PLwJRxp3blEvaxmHgI2iOzNP6KGLSyd4dz&index=55&t=0s

The "Rubin model" by Rubin himself: http://www.stat.columbia.edu/~cook/qr33.pdf

Great introductory course on causality (though, no regression yet): https://www.coursera.org/learn/crash-course-in-causality

Share Cite Edit Follow Flag

edited Dec 16, 2019 at 11:37

answered Dec 16, 2019 at 11:29

Alfred Beit
31 4

Good points. Welcome to CV. – Neil G Dec 16, 2019 at 12:33

My understanding (I'm a causality beginner) is the following:

0

- Linear regression implies causality if your covariates are from a controlled experiment, and your experiment isolates the hypothesized causal factor well (see Linear regression and causality in a randomized controlled experiment).
- Alternatively, (updated thanks to comments), many violations of causality lead to $E(\epsilon|X) \neq 0$. Note that $E(\epsilon|X) \neq 0$ means that we can't draw causal conclusions, but $E(\epsilon|X) = 0$ doesn't mean that we can.

Note that we can't test whether $E(\epsilon|X) = 0$, and there is some circularity in the arguments here.

Share Cite Edit Follow Flag edited Mar 3, 2019 at 22:04 answered Mar 3, 2019 at 6:11

mlstudent
472 3 12

2 Could you elaborate on how $E(\epsilon|X) = 0$ implies causation? – Sextus Empiricus Mar 3, 2019 at 8:36

See this for a detailed discussion stats.stackexchange.com/questions/59588/..., with some nice points made. – mlstudent Mar 3, 2019 at 17:20

could you be a bit more direct. I see no prrof or explanation how or why $E(\epsilon|X) = 0$ implies causation. – Sextus Empiricus Mar 3, 2019 at 18:13

I'm a bit new to causality, but as I understand it there are three major concerns that could make $y = \alpha + \beta x + \epsilon$ not imply causality. One is if there is some other omitted variable causing y, another is if there is an omitted variable causing x, and finally a third is that y may cause x. All will lead to violations of the exogeneity condition. I don't have the math for exactly why but will actually look this up/try to derive it. – mlstudent Mar 3, 2019 at 18:59

A simple counter example. When you generate data $Y \sim N(\mu_Y, \sigma_Y)$ and $X|Y \sim N(a + bY, \sigma_X)$ then you still have $E(\epsilon|X) = 0$ (X and Y are jointly normal distributed). – Sextus Empiricus Mar 3, 2019 at 20:00