The Regression Fallacy

by Potluri Rao In Seattle

Regression equations are the mainstay of economic model building. A study of the failures of economic policies of the major industrial economies suggests that regression equations are misinterpreted to justify misguided political policies.

Regression equations are mathematically unbiased. People who interpret the results can lack basic scientific skills. They are generally interested in job security, not the Truth.

To illustrate the regression fallacy, let us consider an economy with only the variables: x_1, x_2, x_3, x_4 . The variable x_1 is the target, and x_2 is used as a policy tool. The x_3 and x_4 are the other factors that are also impacted by the policy tool x_2, and they in turn impact x_1. The factors x_2, x_3, x_4 directly or indirectly impact the target x_1. The net effect could result in the exact opposite of the intended goal by the controller.

The Yule Notation

Professor Udny Yule invented a notation in 1897 to interpret regression coefficients correctly. Unfortunately, it was ignored by the people who want to abuse regressions to peddle pet policy recommendations.

The Yule notation is actually derivative mathematical functions expressed in a simple minded and easy to understand format.

Consider the regression equation: $x_1 = b_{12} x_2$. The coefficient function b_{12}, estimated from the data, is in the Yule notation. It says, the equation has only two variables: x_1 on the left side, and x_2 on the right side; and b_{12} is a function $f(x_1, x_2) = \sum x_1 x_2 / \sum x_2^2$. Note that b_{12} is a function, not a constant.

In the Yule notation, b_{21} is the coefficient function of a different equation with only two variables: x_2 on the left side, and x_1 on the right side. Thus, $x_2 = b_{21} x_1$.

Notice that $b_{12} = \frac{dx_1}{dx_2}$ and $b_{21} = \frac{dx_2}{dx_1}$.

The Yule notation (b_{12}, b_{21}) states the coefficient functions (derivatives) in a format that is easy to understand by a layman.

To interpret regression coefficient functions correctly we should always think of them as derivative functions. It is the proper scientific approach.

Let us extend the Yule notation to a case of three variables.

$$x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$$

The coefficient function $b_{12.3}$ means the equation has x_1 on the left side, it belongs to x_2 on the right side, and the equation has one other variable x_3; a total of two variables on the right side. Similarly, the coefficient function $b_{13.2}$ means the equation has x_1 on the left side, it belongs to x_3 , and there is one other variable x_2 on the right side.

In the mathematical notation the coefficient functions are partial derivatives:

$$b_{12.3} = \frac{\partial x_1}{\partial x_2}|x_3 \quad \text{and} \quad b_{13.2} = \frac{\partial x_1}{\partial x_3}|x_2.$$

A coefficient with a dot is a partial derivative (∂), and the variables after the dot are held constant. A coefficient without a dot is a total derivative (d).

We may easily generalize this notation as:

$$b_{12.34} = \frac{\partial x_1}{\partial x_2}|x_3 x_4 \quad , \quad b_{13.24} = \frac{\partial x_1}{\partial x_3}|x_2 x_4 \quad , \text{ and } \quad b_{14.23} = \frac{\partial x_1}{\partial x_4}|x_2 x_3$$

to yield the regression equation,

$$x_1 = b_{12.34}\, x_2 + b_{13.24}\, x_3 + b_{14.23}\, x_4.$$

The coefficients are distinct partial derivatives, and should be interpreted as such. They should always be treated as derivatives, not as some arbitrary constants. Each derivative is a different logical question. The inability to distinguish them results in the Regression Fallacy, and abuse of regression analysis.

We know from elementary calculus that in a three variable equation $x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3$,

$$\frac{dx_1}{dx_2} = \frac{\partial x_1}{\partial x_2} \cdot \frac{dx_2}{dx_2} + \frac{\partial x_1}{\partial x_3} \cdot \frac{dx_3}{dx_2}$$

which translates in Yule notation as,

$$b_{12} = b_{12.3} . 1 + b_{13.2} . b_{32}.$$

In any given set of data, the coefficient functions ($b_{12}, b_{12.3}, b_{13.2}, b_{32}$) must always satisfy the above formula. It is an identity. It is the calculus. Regression coefficients are derivatives, not constants. Given any three, we can solve for the fourth. They all tell exactly the same story from different angles (derivatives). That is why we should always treat them as derivatives, not as constants. Yule notation is a constant reminder of the exact nature of the derivatives.

Similarly,

$$\frac{dx_1}{dx_3} = \frac{\partial x_1}{\partial x_2} . \frac{dx_2}{dx_3} + \frac{\partial x_1}{\partial x_3} . \frac{dx_3}{dx_3}$$

translates into Yule notation as

$$b_{13} = b_{12.3} . b_{23} + b_{13.2} . 1.$$

In any given set of data, the coefficient functions ($b_{13}, b_{12.3}, b_{23}, b_{13.2}$) must always satisfy the above formula. Given any three derivatives we can solve for the fourth.

The Yule notation is nothing but the basic calculus presented in a format understandable by a layman to help interpret regression coefficients correctly as derivatives.

All the possible regression coefficients, computed from the same set of data, are mathematically interconnected; they all tell exactly the same story as seen from different angles. The Yule notation issues them appropriate badges for easy identification.

Any scholar trained in the true scientific method would concentrate on asking the proper question for inquiry. Unfortunately, the true scientific method is now extinct; the emphasis is on pandering to the peer review.

The Yule notation is now rare, and Regression Analysis has lost its honesty. Regression analysis has become a tool of abuse of reasoning.

The Abuse of Reasoning

The regression coefficients with no dot are the total derivatives (d), and the ones with dot are partial derivatives (∂). All the derivatives are different angles of vision (planes) of the same data. Each derivative is a different question. All derivatives are linked by calculus. All of them tell exactly the same story from different perspectives.

Naturally, the values of coefficient functions will be different. For example, b_{12} is a total derivative (d); $b_{12.3}$ is a partial derivative (∂). They are answers to different questions.

Let us illustrate the difference between b_{12} and $b_{12.3}$ with an example.

Suppose the Federal Reserve wants to influence x_1 by controlling the policy instrument x_2. It increases the value of x_2 by one unit.

Let us look at a hypothetical scenario.

Suppose x_3 is forbidden to change its position by an act of Congress (price controls). Under the hypothetical act of Congress, the impact of the Fed policy is expressed by the partial derivative $b_{12.3}$, because now x_3 is held constant by an act of Congress. The value of x_1 changes by $b_{12.3}$, that is $\frac{\partial x_1}{\partial x_2}|x_3$ a partial derivative.

If the Congress did not impose price controls, the outcome would have been the total derivative b_{12}, that is $\frac{dx_1}{dx_2}$, ignoring x_3.

The total derivative b_{12} and the partial derivative $b_{12.3}$ are computed with two different regression equations. But, the mathematical relation between them $b_{12}=b_{12.3}.1+b_{13.2}.b_{32}$ as dictated by calculus must always hold.

The coefficients are computed with many different regression equations, but all of them are controlled by the same calculus. All coefficients are of equal status. They all tell exactly the same story from different perspectives (derivatives). Arguments that some coefficients are superior to others are plain rubbish.

Naturally, the impact of the Fed policy would be different under the two scenarios: with and without the act of Congress.

Suppose the Supreme Court ruled that the act of Congress was unconstitutional.

The x_3 is now free to change.

Before the Court ruling, x_1 changed by $b_{12.3}$ (partial). After the ruling, x_1 changed some more because now x_3 was allowed to change.

If x_3 changes by b_{32}, as it did in the past, its impact on x_1 would be: the partial derivative $b_{13.2}$ multiplied by the total derivative b_{32} ; that is $b_{13.2}.b_{32}$.

The total impact on x_1 as a result of a unit change in x_2 would be: $b_{12.3}$ before the Court ruling plus $b_{13.2}.b_{32}$ after the ruling, for a total of $b_{12.3}+b_{13.2}.b_{32}$. It is exactly the same value as the total derivative b_{12}, as shown above. It is elementary calculus.

The coefficients ($b_{12}, b_{12.3}, b_{13.2}, b_{32}$), computed from the same set of data, with three different regression equations, must always satisfy the formula $b_{12}=b_{12.3}.1+b_{13.2}.b_{32}$ as required by calculus. The coefficients are derivatives, not arbitrary constants.

Which is the question we should be asking: Do we want b_{12} the total derivative without the Act of Congress, or $b_{12.3}$ the partial derivative with the Act?

The total derivative is given by the equation $x_1=b_{12}x_2$, and the partial derivative is given by the equation $x_1=b_{12.3}x_2+b_{13.2}x_3$. Which of the two equations should we compute?

The Yule notation makes the equations transparent. One is under the act of Congress (partial) and the other without the act of Congress (total).

Without the Yule notation to guide, researchers don't have the foggiest notion of what question they are asking and why. They are blind people chasing fancy metrics, to get their papers published in journals edited by other blind people, to join the crowd.

The regression equation that produced the $b_{12.3}$ has a positive value for $b_{12.3}$ just as the advocates of Fed interference wanted, with a large t-value of 4.0 and R^2 of 0.99 out of maximum possible 1.0. The regression equation that produced the b_{12} has a negative value for b_{12} with a low t-value of less than 1 and a dismal R^2 of 0.20.

Naturally, the Fed policy advocates want to publish the regression equation with the $b_{12.3}$ which has the right sign, right value, right t-value, and right R^2. They deliberately badmouth the equation with the b_{12} for its wrong sign, wrong value, wrong t-value, and wrong R^2. They tote with great fanfare in the academic publications and social media that the Fed policy would have a positive impact as proven by the data. They are blissfully ignorant of the fact that the regression they selected to promote is the wrong question. They lack the basic scientific skills to ask the right question.

The Fed policy advocates do not understand that from a mathematical perspective b_{12} with the wrong sign and wrong t-value is telling exactly the same story as $b_{12.3}$ with the right sign and right t-value. b_{12} is the correct answer to the right question (without the

Act). $b_{12.3}$ is the correct answer to the wrong question (with the Act). The policy would have a negative impact, unless the Congress steps in and imposes price controls.

Regression analysis always gives the correct answer to the question asked. If you ask a stupid question, you still get the correct answer to your stupid question. Unfortunately, most literature published in the academic journals these days don't have the foggiest notion of what question they are asking.

Regression analysis has become the most abused tool in the hands of policy advocates to promote political agendas.

The question the policy maker should ask is: b_{12} the total derivative, not $b_{12.3}$ the partial derivative. The fancy footwork of metrics (t-value and R^2) is the smokescreen to cover the fact they don't have the basic scientific skills to ask the right question.

Statistical theorems are based on the premise that the question asked is the right question. Statistics has no way of telling that you asked a stupid question. Only the Yule notation (derivative functions) call tell that you asked a stupid question.

Which of the two ($b_{12}, b_{12.3}$) is the stupid question? Only the Yule notation can tell, not statistics. Unfortunately, no one teaches the Yule notation anymore.

Articulate the question to be answered, in the Yule notation, as a derivative function, to make the question transparent for everyone to see. It is the only way to find out if you are asking a stupid question.

In the true scientific method the proper question is: Are you asking the right question?

If you tell me what value you are looking for, I can always find a regression equation with the answer you want from any set of data. There are an infinite number of derivatives to choose from ($b_{12}, b_{12.3}, b_{12.4}, b_{12.34}, b_{12.5}, b_{12.35}, b_{12.45}, b_{12.345}, .. \infty$). All I have to do is add and remove variables until the value that you want shows up. Needless to say, it is a stupid question. Abuse of Reasoning has become a profession.

Reading material Home