;

STATISTICAL HORIZONS

Prediction vs. Causation in Regression Analysis

JULY 8, 2014 BY PAUL ALLISON

In the first chapter of my 1999 book *Multiple Regression*, I wrote

"There are two main uses of multiple regression: prediction and causal analysis. In a prediction study, the goal is to develop a formula for making predictions about the dependent variable, based on the observed values of the independent variables….In a causal analysis, the independent variables are regarded as causes of the dependent variable. The aim of the study is to determine whether a particular independent variable *really* affects the dependent variable, and to estimate the magnitude of that effect, if any."

As in most regression textbooks, I then proceeded to devote the bulk of the book to issues related to causal inference—because that's how most academic researchers use regression most of the time.

Outside of academia, however, regression (in all its forms) is primarily used for prediction. And with the rise of Big Data, predictive regression modeling has undergone explosive growth in the last decade. It's important, then, to ask whether our current ways of teaching regression methods really meet the needs of those who primarily use those methods for developing predictive models.

Despite the fact that regression can be used for both causal inference and prediction, it turns out that there are some important differences in how the methodology is used, or should be used, in the two kinds of application. I've been thinking about these differences lately, and I'd like to share a few that strike me as being particularly salient. I invite readers of this post to suggest others as well.

1. *Omitted variables*. For causal inference, a major goal is to get unbiased estimates of the regression coefficients. And for non-experimental data, the most important threat to that goal is omitted variable bias. In particular, we need to worry about variables that both affect the dependent variable and are correlated with the variables that are currently in the model. Omission of such variables can totally invalidate our conclusions.

With predictive modeling, however, omitted variable bias is much less of an issue. The goal is to get optimal predictions based on a linear combination of whatever variables are available. There is simply no sense in which we are trying to get optimal estimates of "true" coefficients. Omitted variables are a concern only insofar as we might be able to improve predictions by including variables that are not currently available. But that has nothing to do with bias of the coefficients.

2. R^2. Everyone would rather have a big R^2 than a small R^2, but that criterion is more important in a predictive study. Even with a low R^2, you can do a good job of testing hypotheses about the effects of the variables of interest. That's because, for parameter estimation and hypothesis testing, a low R^2 can be counterbalanced by a large sample size.

For predictive modeling, on the other hand, maximization of R^2 is crucial. Technically, the more important criterion is the standard error of prediction, which depends both on the R^2 and the variance of y in the population. In any case, large sample sizes cannot compensate for models that are lacking in predictive power.

3. *Multicollinearity*. In causal inference, multicollinearity is often a major concern. The problem is that when two or more variables are highly correlated, it can be very difficult to get reliable estimates of the coefficients for each one of them, controlling for the others. And since the goal is accurate coefficient estimates, this can be devastating.

In predictive studies, because we don't care about the individual coefficients, we can tolerate a good deal more multicollinearity. Even if two variables are highly correlated, it can be worth including both of them if each one contributes significantly to the predictive power of the model.

4. *Missing data*. Over the last 30 years, there have been major developments in our ability to handle missing data, including methods such as multiple imputation, maximum likelihood, and inverse probability weighting. But all these advances have focused on parameter estimation and hypothesis testing. They have not addressed the special needs of those who do predictive modeling.

There are two main issues in predictive applications. First, the fact that a data value is missing may itself provide useful information for prediction. And second, it's often the case that data are missing not only for the "training" sample, but also for new cases for which predictions are needed. It does no good to have optimal estimates of coefficients when you don't have the corresponding *x* values by which to multiply them.

Both of these problems are addressed by the well-known "dummy variable adjustment" method, described in my book *Missing Data*, even though that method is known to produce biased parameter estimates. There may well be better methods, but the only article I've seen that seriously addresses these issues is a 1998 unpublished paper by Warren Sarle.

5. *Measurement error*. It's well known that measurement error in predictors leads to bias in estimates of regression coefficients. Is this a problem for a predictive analysis? Well, it's certainly true that poor measurement of predictors is likely to degrade their predictive power. So efforts to improve measurement could have a payoff. Most predictive modelers don't have that luxury, however. They have to work with what they've got. And after-the-fact corrections for measurement error (e.g., via errors-in-variables models or structural equation models) will probably not help at all.

I'm sure this list of differences is not exhaustive. If you think of others, please add a comment. One could argue that, in the long run, a correct causal model is likely to be a better basis for prediction than one based on a linear combination of whatever variables happen to be available. It's plausible that correct causal models would be more stable over time and across different populations, compared with ad hoc predictive models. But those who do predictive modeling can't wait for the long run. They need predictions here and now, and they must do the best with what they have.

Tweet

Comments (58)

58 RESPONSES

> Andy says:
> July 17, 2014 at 10:48 am
> Hi Dr. Allison,
> I have had some colleagues mention to me lately that if the sample size is large enough one can completely ignore multicollinearity and can conduct inference on the associated coefficients with no concerns. Primarily they reference a 2001 chapter by Kent Leahy in the 'data mining cookbook' but our own readings of that source seem to suggest otherwise.
>
> Traditional statistical thought on the topic says that this statement is only true when you have properly specified the model and no relevant predictors are missing. In practice, the only thing we know for sure about ALL our models in a business setting is that they are NOT specified properly and exclude something that matters, either because we can't legally consider it, or it's something that we can't measure like the consumer's attitude the day of the event.
>
> We've done some simulations and what we found is that it does indeed hold true if you have correctly specified the model, but that it breaks down when you omit an element of the true model and include correlated proxies instead.
>
> How would you respond to the absolute claim that "if n is large enough, you can completely ignore multicollinearity and interpret coefficients without concern?"
>
> I'm hopeful your thoughts on the specific matter of multicollinearity and inference on the betas in the presence of multicollinearity can help bring these discussions to conclusion.
>
> Reply

Paul Allison says:
July 17, 2014 at 4:11 pm

I agree that large n should not alleviate concerns about multicollinearity. When variables are collinear, very small changes in the model specification can have big effects on the results. So multicollinear data are not very robust to specification errors.

Reply

Kent Leahy says:
January 21, 2015 at 7:10 pm

The article by Kent Leahy re increasing the sample size as a correction for multicollinearity is only valid for models developed for predictive purposes, as the article makes clear. In other words , “model specification” is not a concern under such circumstances.

Reply

Ming says:
July 30, 2014 at 1:36 pm

Dear Dr. Allison: another difference between the two is use of link function. Wnen the dependent variable is a rate with values limited to 0 to 1, link function or transformtion is usually recommended for making the distribution closer to some well-known distributions as to mitigate estimation bias. For predictive model, I found models without use of link function or transformation usually perform better than otherwise, bacause error in estimation are usually magnified by inversion of the transformation. Can you share your thougt on this topic?

Reply

Paul Allison says:
July 30, 2014 at 2:15 pm

Well, I would think you would want your predictions limited to the 0-1 interval, which is one of the main reasons for using, say, a logit or probit link.

Reply

Jacob Arendt says:
August 5, 2014 at 2:21 am

Highly interesting topic. You state that one would samt a high r2 in predictive models to minimize within sample prediction error. But wouldn’t be even better to look at out-of-sample ? I guess it boils Down to assumptions about similarities in distributions of samples (within sample and prediction sample) whether these are different? Do you agree? Best Jacob

Reply

Paul Allison says:
August 5, 2014 at 9:03 am

I agree that the assessment out-of-sample prediction is much more important in predictive modeling than in causal inference.

Reply

Ashutosh Tamhane says:

October 23, 2014 at 4:32 pm

In causal modeling, focus is on including variables that qualify as “confounders” for the exposure(independent variable of interest)-outcome association. While in predictive modeling, the variables included may not necessarily be qualified as “confounders”. Your opinion please.

Reply

Paul Allison says:

October 27, 2014 at 6:46 am

Agreed.

Reply

Syu says:

December 6, 2014 at 5:41 am

why we can not one independent variable cause the change in the dependent variable in multiple regression? Which means why we can not say causation in multiple regression?

Reply

Paul Allison says:

December 6, 2014 at 1:11 pm

Sorry, but I don’t understand this question.

Reply

Bilal Khan says:

December 27, 2014 at 4:58 pm

Interesting post. I just have few questions related to predictive modelling versus causal modeling. First of all, in causal modeling controlling for variables that are the effect of treatment variable will lead to of estimation bias. So we need to control for pre-treatment and variables not effected by our treatment variable. In predictive modeling, is post estimation bias a problem too. Can we control for effect of treatment variable in prediction models like propensity score matching or doubly robust regression where causality is based on outcome and treatment models as good predictive models. Moreover, in presence of multicollinearity or even small multicollinearity, due to shared variance, some coefficients in the model may be counter intuitive such as wrong signs or becoming insignificant. Thus, I used decomposition of R square to discern their relative importance instead of standardized beta weights. It showed almost 15 percent contribution of a variable which had become insignificant. Should we trust 15 percent variation (Shapely value regression model) or the insignificant standardized beta weights

Reply

Paul Allison says:

December 29, 2014 at 10:58 am

In predictive modeling, controlling for a variable that is affected by a “treatment” variable should not be a cause for concern.
I don’t fully understand your question about propensity score matching.
There’s usually not a lot of difference between standardized beta weights and decomposition of R square.

Reply

heba says:

January 29, 2015 at 10:13 am

Dear Dr Paul,

Their is an argument that we can not use regression for causation ? what do think

if correlation does not imply causation and regression too , so what test can imply causation?

Thank you

Reply

Paul Allison says:
January 29, 2015 at 10:23 am

The gold standard is a randomized experiment. But regression can be helpful in ruling out alternative hypotheses.

Reply

Terri says:
February 3, 2015 at 5:24 am

Dear Dr Allison,

A very helpful article – thank you.

On variable selection for a predictive model and collinearity: one approach (given a large sample and enough events) is to include all available variables (assuming less than, say, 20). However, given that we want as precise a prediction as possible, should we be checking not to include variables that are associated with another variable but not with the outcome, on the grounds that such variables widen the standard errors of various coefficients, and even if we are not primarily interested in the coefficients themselves, their lack of precision will feed through to lack of precision in the prediction. If there is anything to be said for this argument, then would it not also apply to avoiding collinearity in a predictive model?

Separately, are we not in practice usually also still interested in the coefficients? That is to say, having identified someone for whom we predict a high risk of death, we then want to intervene so we would look to the coefficients to tell us about the relative risks of different factors so that we can tell someone how to reduce their risk.

Many thanks for any thoughts,
Terri

Reply

Paul Allison says:
February 3, 2015 at 6:40 am

I agree with your first point. As to the second, if your goal is to reduce risk, then you're moving into the realm of causality and different considerations apply. But there are many applied situations where intervention is not the goal.

Reply

Nate says:
September 1, 2015 at 3:05 am

Dear Dr Allison,

Thank you for the enlightening post.

I just had a question regarding variable selection when building a predictive model. In such a situation where the predictive variable in question is someone associated with the outcome in question, would it still be considered logical to include it in the analysis? Say, for example, the inclusion of the predictive variable serum creatinine levels in a model to predict risk of progression to renal failure (itself characterized in the data by the use of serum creatinine parameters)

Many Thanks!

Nate

Reply

Paul Allison says:
September 3, 2015 at 4:21 pm
Well, if the main goal is prediction, I don't see a problem. That presumes that, when using the model, one would have knowledge of serum creatinine levels before knowing whether the patient will progress to renal failure.

Reply

Ambarish Dutta says:
November 24, 2015 at 1:01 am
Excellent article. Can I ask a question which may not be directly relevant to this causal vs predictive dichotomy discourse? Is multivariable analysis as robust a technique as matching to get unconfounded estimates, especially when different units in a community trial (unfortunately was not randomized or matched at the intervention and data collection stage) are being compared?

Reply

Paul Allison says:
December 10, 2015 at 4:25 pm
This is controversial. Proponents of propensity score matching claim that it is more robust to situations in which there is not much overlap between treatment groups in the distributions of covariates. However, unless the pool of potential matches is large, matching can run into problems with poor matches or an insufficient number of matches. It is also not well suited to quantitative "treatments" and not well developed for categorical treatments with multiple categories.

Reply

Jim Grace says:
December 27, 2015 at 3:22 am
Dear Dr. Alison,

Thank you for an excellent post. Two questions if I may,

(1) Do you know of any published, more extensive treatments of the dichotomy between prediction-only and causal modeling philosophy (I know of plenty that are one or the other)?

(2) Burnham and Anderson (2004, Soc. Methods & Res. 33:261-304) argue strongly for model averaging, repeatedly saying "When prediction is the goal". Their arguments are all fine for that limited sphere of interest. As a causal modeler (SEM primarily), I have no problem using multimodel inference with a set of causal models, but find the concept of "model averaging" out of sync with my ideas about how to critique causal models. Any thoughts on that?

Reply

Paul Allison says:
December 28, 2015 at 7:08 am
(1) No, I don't.
(2) I'm also a bit skeptical of model averaging for causal inference. But then I haven't really carefully evaluated the arguments pro and con.

Reply

David Sabaj-Stahl says:
January 8, 2016 at 6:31 pm

"R2. Everyone would rather have a big R2 than a small R2, but that criterion is more important in a predictive study. Even with a low R2, you can do a good job of testing hypotheses about the effects of the variables of interest. That's because, for parameter estimation and hypothesis testing, a low R2 can be counterbalanced by a large sample size."

Generally I agree with your assessment of large v. small r2 values. More data are usually better, but I've read that a very large dataset can generate artificially small p values. In other words, the p may appear to indicate a significant relationship in conjunction with a small r2, but this could be an artifact caused by a very large dataset.

I am curious about your opinions, as I may have observed the effect in some of my data. Thanks!

Reply

Paul Allison says:
January 25, 2016 at 10:32 am

It's certainly true that with large samples, even small effect sizes can have low p-values. This isn't an "artifact" in itself, but it does mean that small biases in coefficients can yield statistically significant results. So with large samples, you need to evaluate the magnitude of an effect, not just its statistical significance. Alternatively, focus on confidence intervals rather than p-values.

Reply

Colin Vance says:
February 3, 2016 at 10:31 am

Thanks for this excellent post. I have a question concerning multicollinearity, which you say is a major concern in causal analysis. Specifically, you note that since the goal is accurate coefficient estimates, high correlation between two variables can be devastating because low precision on the coefficient estimates of the variables may result. But my understanding is that accuracy is not compromised by multicollinearity; the OLS estimate remains unbiased. Conversely, were you to omit one of the correlated variables, precision would surely increase, but accuracy would be lost. This would seem to be a greater danger, false confidence in a biased estimate. Is it not better to accept multicollinearity as a cost of unbiasedness if causal analysis is the main aim?

Reply

Paul Allison says:
February 8, 2016 at 8:27 am

I agree.

Reply

Dan says:
August 22, 2019 at 3:08 pm

The OLS estimate is unbiased but unreliable because of the high variance.
Penalization such as in ridge regression will reduce the total variance but at the price of bias.
The problem is to balance the two.

Reply

Hammond thurib says:

March 8, 2016 at 4:14 pm

Dear Dr. Allison,

Thank you for this clarifying article. I come from a machine learning background and have entered the field of epidemiology. From predictive modelling I find the methods for model validation very useful e.g., for avoiding overfitting using cross-validation and training/test sets. In causal modelling I don't see method validation in the same fashion and wonder why this is the case? Is overfitting and overestimation of associations not and issue in causal analysis?

Best,

Hammond

Reply

> Paul Allison says:
>
> March 9, 2016 at 12:53 pm
>
> I definitely think that issues regarding overfitting and cross-validation should be more widely addressed in causal modeling. Why aren't they? Here are couple possible reasons: 1. Causal modelers typically work with smaller sample sizes and are, therefore, reluctant to split up their data sets. 2. Causal modelers don't actually have to address the issue of how well their models can perform in a new setting.
>
> Reply

Zebidiah says:

June 6, 2016 at 2:39 pm

Thank you Dr. Allison. I'm convinced that there are 3 types of people in this world. those that are good at math and those that aren't.

Reply

Lili says:

September 6, 2016 at 9:26 pm

Dear Dr. Allison,

This post is very interesting. I was wondering whether you have published a formal article in a 'formal' journal that I could cite regarding those important differences in methodology between prediction and causal multiple regression analyses. Thank you!

Reply

> Paul Allison says:
>
> September 21, 2016 at 9:36 am
>
> No, I have not published an article on this topic.
>
> Reply
>
> > Laura Dee says:
> >
> > July 10, 2018 at 2:11 pm
> >
> > Dear Dr. Allison,
> >
> > I am replying to this post to see if you or others now have a publication that formally lays out these important distinctions. Thanks!
> >
> > Laura
> >
> > Reply

Giuseppe says:

March 30, 2017 at 6:06 am

Dear Dr Allison,
a very interesting article – thank you.
I totally agree with you.
Moreover, if the focus is about prediction, let me add:
– about measurement error I have a more radical view. I think that if proxy variable, in term of fitting (and out of sample statisctics) are better of the original one … then proxy variable is simply better than original.
– You have not talked about simultaneity. This topic is very important about causal inference (reverse causation problem) but in term of prediction … it is not an issue.
More in general, even if many textbooks are not clear about this poit, it seems me that in "prediction world" … endogeneity problem at all is definitely not an issue.
Probably the overfitting is a main issue but out off sample test help us about this.

Are you agree ?

Reply

Tran Huu Bich says:

June 13, 2017 at 12:37 am

Thank you, Dr. Allison for opening this important and interesting topic. I have been looking for this topic and found it. I think you would better to publish this article in the open access for colleagues to learn more about this issue.

Reply

Joe says:

September 6, 2017 at 6:34 pm

Does that mean if one would like to build predictive model, spurious regression is fine if only if it could provide good prediction result?

Reply

Paul Allison says:

September 8, 2017 at 9:43 am

Yes

Reply

liza says:

September 27, 2017 at 6:27 am

Thank you so much for this post! I was learning regression in different courses, and it always confused me, since some were addressing it as predictive and some as causal, and the differences was never discussed. Only after reading your post, everything now makes better sense.

Reply

Veronica says:

April 10, 2018 at 5:01 pm

Hello, thanks for this posting!
When discussing the predictive and/or causal value of the multiple regression, what is the relevance of having cross sectional or longitudinal data?
For example, if one wants to test causality, I understand that having data from different time points should be a must. Am I right?
Then, are predictive models more suitable for cross sectional data?
Thanks!

Reply

Paul Allison says:

May 25, 2018 at 12:54 pm

1. Longitudinal data are desirable for making causal inferences but they are no panacea.
2. There are situations in which cross-sectional data can be adequate. If you know from theory or just common sense that Y cannot affect X, then cross-sectional data may be adequate.

Reply

Simon Balthazer says:

April 24, 2018 at 4:14 am

Can I ask a somewhat related but different question:

What is the difference – between an explanatory variable (i.e. key independent variable of interest) and control variables? Mathematically, are they not treated equally as X1, X2,…Xn? And is the only difference in our interpretation of their beta-coefficients (or log-odds, as the model may be)?

Reply

Paul Allison says:

May 25, 2018 at 12:37 pm

Mathematically there's no difference. It's all about what we care about and what we don't care about.

Reply

Kim says:

June 21, 2018 at 4:21 am

Thank you very much for this post!

Question: I am trying to run a (weighted) binary logit regression with personal characteristics as independent variables using a large survey data. The dependent variable may be considered a rare event given that only 2% of the sample have Y=1.

In this case, I am trying to predict a person's probability of Y=1 given his/her characteristics.
Is this what you would consider "predictive modeling"? – and as such, omitted variables are not as much of an issue?

I am wondering because I am running diagnostic tests after the weighted logit and get a McFadden R^2 above 0.2 (0.2 – 0.4 suggests an "excellent fit") but linktest suggests mis-specification (significant _hatsq).

Thank you very much!

Reply

Paul Allison says:

June 21, 2018 at 9:39 am

Sounds like predictive modeling to me. And for that, you do the best you can with the variables you have. Also, getting an R^2 of .2 with only 2% of the cases having events is pretty good. But the linktest suggests that you might do a little bit better with a different link function, or with some transformation of the predictors.

Reply

Peter says:

October 30, 2018 at 8:39 am

One difference that is worth noting is that the predictive model can be stated in terms of conditional distributions: E(Y|X) = beta*X. In this context, the strict exogeneity assumption used routinely by econometricians is superfluous, as it is automatically satisfied. It would be helpful if econometricians would more often clarify which model they are talking about, and which assumptions are needed for each.

Reply

Mari Palta says:

February 2, 2019 at 11:59 am

In glancing through this long sequence I do not see one thing addressed that I have been wondering about: it seems that a predictive model that favors causative factors would be more portable to new settings. This aspect does not seem to be solved by validation samples from the same setting. Any work on this?

Reply

Paul Allison says:

February 26, 2019 at 2:43 pm

I definitely agree that, in principle, models that capture the correct causal relationship should be the most generalizable to new settings. I am not aware of any work on this, but that doesn't mean there isn't something out there. It would be difficult to research this in any general way, however, because every substantive application will be different.

Reply

Dan says:

August 22, 2019 at 8:34 pm

Here is another difference:

Regularization, e.g., ridge regression, is needed for both but for different reasons. In inference we need regularization to temper the volatility of estimates when the data is multicollinear and in prediction we need it to temper over fitting. The computation of the hyper parameter(s) is also different. In inference, for example, sometimes the L-curve is used or the trace of the coefficients, etc. but for prediction it is cross validation.

Reply

Daniel P Vasilaky says:

September 4, 2019 at 2:34 pm

This paper may shed additional light on the subject:

To Explain or to Predict?

https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf

Reply

Paul Allison says:

September 5, 2019 at 2:40 pm

Great suggestion! I haven't had a chance to carefully read this article, but it looks excellent.

Reply

Selva says:

December 27, 2019 at 8:51 am

Hi Dr.Allison,

Thanks for this post. Very much useful.

I have few questions. can you help me with these?

1) I am working on a predictive model for healthcare related application (disease prediction). being a person from non-healthcare domain, what I am trying to do is build a predictive model based on algorithms like Random Forests, Boosting (tree based model) etc which can help me know the combination of features that can help me predict the outcome. But is it to possible to add causal model ability to this? refer below

2) What I mean is, will I be able to build a causal model on my dataset and identify the important features/columns that are significant and use them to build the predictive model. Is this even right?

3) Next, if I have to build a causal model, I read up online that in Logistic regression, we have to adjust for confounding variables. Once I adjust for confounding variables and get the list of significant variables, I can ten use them in predictive model? Is my thought process right?

4) How can I adjust confounders in logistic regression? Currently when I use `python` statsmodel approach, it doesn't consider confounders. Can you know let me know how can we do this confounding adjustment programatically?

Reply

Paul Allison says:
December 27, 2019 at 11:48 am

1. In principle, yes. But some techniques, like logistic regression, are more suitable for causal modeling while others, like random forests, not so much. And there are different considerations in building a causal model as opposed to a predictive model.

2. You could use random forests to suggest variables/features that should go into your causal model.

3. A lot of careful thought needs to go into a causal model. You might want to check out Stephen Morgan's book, Counterfactuals and Causal Inference.

4. In logistic regression, there's no operational distinction between causal variables and confounders. They're all just predictor variables in the equation.

Reply

Jordi says:
February 20, 2020 at 8:44 pm

Thanks for the interesting post I just stumbled across.

Another reference for those interested in some further reading is contained in the last section of the following Science article:

https://science.sciencemag.org/content/sci/346/6210/1243089.full.pdf

Reply

Bert Breitenfelder says:
December 23, 2020 at 5:10 am

Thank you for your post.

Would it make sense to add simultaneity to your list? By this I mean the problem when two variables are co-determined, with each affecting the other as in
y = a1*x + a2*z + u
x = b1*y + b2*z + v.

This simultaneity makes estimating the coefficients difficult but reading your post, it seems to me that this is less of a problem when we are trying to predict x and y.

Reply

Paul Allison says:
December 23, 2020 at 8:52 am

I agree. Not a serious problem for prediction.

Reply

Anvr says:
January 29, 2021 at 11:13 am

Dear Paul,

Thank you for this wonderful resource. I want to ask a question. Two variables are correlated to each other. We hypothesize that one should predict or cause other. If the relationship between IV and DV is insignificant in regression analysis, can introduction of a moderating variable turn this insignificant relationship to significant relationship. For MGA, sometimes groups like gender turn this insignificant relation ship to significant for one group, but I am concerned with continuous moderating variables. What I assume, if there is no significant relationship between IV and DV, if moderator variable turns this relationship to significant one, it means that moderator is causing IV and becomes parent of IV and IV is now causing DV and becomes parent of DV. This turns the whole scenario to opposite, as now the model becomes mediating one. Please send me your opinion on this. I have seen few papers published in good journals that report similar situation. According to definition of moderation process it is not possible for a moderator to cause significant relationship between IV and DV.

Reply

Paul Allison says:
February 8, 2021 at 9:33 am

Yes, introduction of a moderating variable can cause a non-significant relationship to become significant. Let's take your gender example. Suppose that X has a strong positive effect on Y among males, and a strong negative effect among females. If you do a simple bivariate regression of Y on X, the two effects can cancel out, leading to a non-significant effect. But if you introduce the interaction of X with gender, you see strong evidence for the separate effects. But the same thing can happen with a continuous moderator Z. For some values of Z, the effect of X on Y is positive and for other values of Z the effect of X on Y is negative. Without the interaction, the effects may cancel out. No need to suppose that Z is causing X, and it's not turning moderation into mediation.

Reply

LEAVE A REPLY

Name:*

E-mail (will not be published):*

Comment:

SUBMIT COMMENT

About | Resources | FAQs | Seminars | Instructors | Code Horizons | Blog | Contact Us .

Connect: