Prediction vs. Causation in Regression Analysis

JULY 8, 2014 BY PAUL ALLISON

In the first chapter of my 1999 book *[Multiple Regression](http://www.amazon.com/Multiple-Regression-Undergraduate-Research-Statistics/dp/0761985336/ref=la_B001H6KWN6_1_3?s=books&ie=UTF8&qid=1404836455&sr=1-3" \t "_blank)*, 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](http://www.amazon.com/Missing-Quantitative-Applications-Social-Sciences/dp/0761916725/ref=la_B001H6KWN6_1_5?s=books&ie=UTF8&qid=1404836455&sr=1-5" \t "_blank)*, 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](http://www.academia.edu/3337303/Prediction_with_missing_inputs" \t "_blank)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.