Journal of Econometrics 237 (2023) 105516

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

What is a standard error?☆

Andrew Gelman

Department of Statistics and Department of Political Science, Columbia University, New York, United States of America

ARTICLE INFO

Article history:
Received 7 August 2023
Received in revised form 7 August 2023
Accepted 13 August 2023
Available online 7 September 2023

Overview

In statistics, the standard error has a clear technical definition: it is the estimated standard deviation of a parameter estimate. In practice, though, challenges arise when we go beyond the simple balls-in-urn model to consider generalizations beyond the population from which the data were sampled. This is important because generalization is nearly always the goal of quantitative studies. In this brief paper we consider three examples.

What is the standard error when the bias is unknown and changing (my bathroom scale)?

I recently bought a cheap bathroom scale. I took the scale home and zeroed it—there's a little gear in front to turn. I tapped my foot on the scale, it went to -1 kg, I turned the gear a bit, then it went up to $+2$, then I turned a bit back to get it exactly to zero, and tapped again …it was back at -1. That was frustrating, but I still wanted to estimate my weight. So I got on and off the scale multiple times. The first few measurements were 66 kg, 65.5 kg, 68 kg, and 67 kg. A lot of variation! To get a good estimate in the presence of variation, it is recommended to take multiple measurements. So I did so. After 46 measurements, I got bored and stopped. The resulting measurements had mean 67.1 with standard deviation 0.7, hence a standard error of $0.7/\sqrt{46} = 0.1$.

Would I want to use the resulting 95% confidence interval, 67.1 ± 0.2? Of course not! The whole scale is off by some unknown amount. What, then, to do? One approach would be to calibrate, either using a known object that weighs in the neighborhood of 67 kg or else my own weight measured on an accurate instrument. If that is not possible, then I would want a wider uncertainty interval to account for the uncertainty in the scale's bias. The usual purpose of a standard error is to attach uncertainty to an estimate, and for that purpose, the usual standard error formula is inappropriate.

How do you interpret standard errors from a regression fit to the entire population (all 50 states)?

Sometimes we can all agree that if you have a whole population, your standard error is zero. This is basic finite population inference from survey sampling theory, if your goal is to estimate the population average or total. Consider a regression fit to data on all 50 states in the United States. This gives you an estimate and a standard error. Maybe the estimated coefficient of interest is only one standard error from zero, so it is not "statistically significant". But what does that mean, if you have the whole population? You might say that the standard error does not matter, but the internal variation in the data still seems relevant, no?

One way to think about this is to imagine the regression being used for prediction. For example, you have all 50 states, but you might use the model to understand these states in a different year. So you can think of the data you have from the

☆ For *Journal of Econometrics*. We thank the U.S. National Science Foundation, National Institutes of Health, and Office of Naval Research for partial support of this work and Serena Ng and Elie Tamer for suggesting this topic. The bathroom scale story is taken from here: https://statmodeling.stat.columbia.edu/2023/01/06/god-is-in-every-leaf-of-every-tree-bathroom-scale-edition/.

E-mail address: ag389@columbia.edu.

https://doi.org/10.1016/j.jeconom.2023.105516

50 states as being a sample from a larger population of state-years. It is not a random or representative sample, though, in that it is data from just one year. So to get the right uncertainty you will need to use a multilevel model or clustered standard errors. With data from only one cluster, some external assumptions will be needed to compute the standard error. Alternatively, one could just use the standard error that pops out of the regression, which would correspond to an implicit model of equal variation between and within years. Just because you have an exhaustive sample, that does not mean that the standard error is undefined or meaningless.

How should we account for nonsampling error when reporting uncertainties (election polls)?

In an analysis of state-level pre-election polls, we have found the standard deviation of empirical errors – the difference between the poll estimate and the election outcome – to be about twice as large as would be expected from the reported standard errors of the individual surveys. This sort of nonsampling error is usual in polling; what is special about election forecasting is that here we can observe the outcome and thus measure the total error directly. The question then arises: what standard error should a pollster report? The usual formula based on sampling balls from an urn (with some correction for weighting or survey adjustment) gives an internal measure of uncertainty but does not address the forecasting question. It would seem better to augment the standard error based on past levels of nonsampling error, but then the question arises of what to do in other sampling settings where no past calibration is available. In election polling we have some sense of that extra uncertainty; it seems wrong to implicitly set it to zero when we do not know what to do about it.

How, then, should we interpret the standard error from textbook formulas or when fitting a regression? We can think of this as a lower-bound standard error or, more precisely, as a measure of variation corresponding to a particular model.

Summary

The appropriate standard error depends not just on the data and sampling model but also on the generalization of interest, and the model of variation across units and over time corresponding to the uses to which the estimate will be put. Deciding on a generalization of interest in a sampling or regression problem is similar to the problem of focusing on a particular average treatment effect in causal inference: thinking seriously about your replications (for the goal of getting the right standard error) and inferential goals, you might well get a better understanding of what you are trying to do with your model.