Regression: What's it all about?[1]

Andrew Gelman[2]

5 Jan 2015

Regression plays three different roles in applied statistics:

1. A specification of the *conditional expectation* of *y* given *x*;
2. A *generative model* of the world;
3. A method for *adjusting* data to generalize from sample to population, or to perform causal inferences.

We could also include *prediction*, but I prefer to see that as a statistical operation that is implied for all three of the goals above: conditional prediction as a generalization of conditional expectation, prediction as the application of a linear model to new cases, and prediction for unobserved cases in the population or for unobserved potential outcomes in a causal inference.

I was thinking about the different faces of regression modeling after being asked to review the new book, *Bayesian and Frequentist Regression Methods*, by Jon Wakefield, a statistician who is known for his work on Bayesian modeling in pharmacology, genetics, and public health.

Wakefield's view of regression is modern in being mostly nonlinear (with two chapters devoted to linear regression and six chapters on nonlinear models). One difficulty when considering nonlinear models is there are so many that it's hard to fit them all in one volume. This book considers generalized linear models and estimating equations but not ordered logit/probit, item-response, or proportional hazard (Cox) regression. There is some material on statistical methods such as kernels and trees that are associated with machine learning, but nothing on wavelets, Gaussian processes, or Bart. Omissions are unavoidable—in this age of Wikipedia, there's no need for a book to contain everything—so this is not a major criticism. But I'd like to hear Wakefield's opinion on these other models, based on his decades of research and applied experience.

The book is theory-heavy (indeed, it is not until around page 200 that the models begin) but at the same time is illustrated with many real examples from Wakefield's research, so I think it will be valuable for a range of audiences. In addition, Wakefield has a website with R code to reproduce every figure and analysis: http://faculty.washington.edu/jonno/regression-methods.html

Now on to Bayesian and frequentist regression. Here is Wakefield's summary:

> For small samples, the Bayesian approach with thoughtfully well-specified priors is often the only way to go because of the difficulty in obtaining well-calibrated frequentist intervals. . . . For medium to large samples, unless there is strong prior information that one wishes to incorporate, a robust frequentist approach . . . is very appealing since consistency is guaranteed under relatively mild conditions. For highly complex models . . . a Bayesian approach is often the most convenient way to formulate the model . . .

All this is reasonable, and I appreciate Wakefield's effort to delineate the scenarios where different approaches are particularly effective. Ultimately, I think that any statistical problem that can be solved Bayesianly can be solved using a frequentist approach as well (if nothing else, you can just take the Bayesian inference and from it construct an "estimator" whose properties can then be studied and perhaps improved) and, conversely, effective non-Bayesian approaches can be

---

[1] Review of *Bayesian and Frequentist Regression Methods*, by Jon Wakefield. To appear in *Statistics in Medicine*.

[2] Department of Statistics and Department of Political Science, Columbia University, New York.

mimicked and sometimes improved by considering them as approximations to posterior inferences. More generally, I think the most important aspect of a statistical method is not what it does with the data but rather what data it uses. That all said, in practice different methods are easier to apply in different problems.

A virtue—and a drawback—of Bayesian inference is that it is all-encompassing. On the plus side, once you have model and data, you can turn the crank, as the saying goes, to get your inference; and, even more importantly, the Bayesian framework allows the inclusion of external information, the "meta-data," as it were, that come with your official dataset. The difficulty, though, is the requirement of setting up this large model. In addition, along with concerns about model misspecification, I think a vital part of Bayesian data analysis is checking fit to data—a particular concern when setting up complex models—and having systematic ways of improving models to address problems that arise.

I would just like to clarify the first sentence of the quote above, which is expressed in such a dry fashion that I fear it will mislead casual or uninformed readers. When Wakefield speaks of "the difficulty in obtaining well-calibrated frequentist intervals," this is not just some technical concern, that nominal 95% intervals will only contain the true value 85% of the time, or whatever. The worry is that, when data are weak and there is strong prior information that is not being used, classical methods can give answers that are not just wrong—that's no dealbreaker, it's accepted in statistics that any method will occasionally give wrong answers—but clearly wrong, obviously wrong. Wrong not just conditional on the unknown parameter, but conditional on the data. Scientifically inappropriate conclusions. That's the meaning of "poor calibration." Even this, in some sense, should not be a problem—after all, if a method gives you a conclusion that you know is wrong, you can just set it aside, right?—but, unfortunately, many users of statistics consider to take $p < 0.05$ or $p < 0.01$ comparisons as "statistically significant" and to use these as motivation to accept their favored alternative hypotheses. This has led to such farces as recent claims in leading psychology journals that various small experiments have demonstrated the existence of extra-sensory perception, or huge correlations between menstrual cycle and voting, and so on.[3]

In delivering this brief rant, I am not trying to say that classical statistical methods should be abandoned or that Bayesian approaches are always better; I'm just expanding on Wakefield's statement to make it clear that this problem of "calibration" is not merely a technical issue; it's a real-life concern about the widespread exaggeration of the strength of evidence from small noisy datasets supporting scientifically implausible claims based on statistical significance.

Frequentist inference has the virtue and drawback of being multi-focal, of having no single overarching principle of inference. From the user's point of view, having multiple principles (unbiasedness, asymptotic efficiency, coverage, etc.) gives more flexibility and, in some settings, more robustness, with the downside being that application of the frequentist approach requires the user to choose a method as well as a model. As with Bayesian methods, this flexibility puts some burden on the user to check model fit to decide where to go when building a regression.

Regression is important enough that it deserves a side-by-side treatment of Bayesian and frequentist approaches. The next step to take the level of care and precision that is taken when considering inference and computation given the model, and apply this same degree of effort to the topics of building, checking, and understanding regressions. There are a number of books on applied regression, but connecting the applied principles to theory is a challenge. A related challenge in exposition is to unify the three goals noted at the beginning of this review. Wakefield's book is an excellent start.

---

[3]Andrew Gelman, "The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective, *Journal of Management* (2014).