DOES THE MODEL MAKE SENSE?

J. MICHAEL STEELE

Abstract. This is a class note that is not for publication — at least anytime soon or in anything like the present form. The intention is to collect the ideas that seem most sensible for deciding if a model makes sense. Please consider contributing to the conversation.

1. A Traditional Method — First Level

Consider the almost infinitely general model:

$$y = F(x, \theta, \epsilon).$$

Here y can be as complicated as you like, say a vector time series or a highly structured table. Similarly, the vector x of covariates can be very complicated, and the same is true for the error term vector ϵ. Naturally, we suppose that you have data (y, x) and that you have the capacity to estimate θ.

Traditional methods of assessing model adequacy tend to focus on the residual vector

$$r = y - F(x, \hat{\theta}, 0)$$

where I have included a non-traditional 0 in my F because F is indeed a function of three vectors, despite the conventional behavior of allowing its definition to morph from line to line.

Moreover, in this notation it is explicit that we permit noise to appear non-additively. One might note that here it would be bad form to have called the residuals $\hat{\epsilon}$. When the epsilons appear non-linearly in the model, the residuals r are **simply residuals** — not ϵ wanabes. We will see shortly that there is another good reason to keep the third slot in plain view.

The traditionalist now looks at the residuals and makes a call. He hopes that the residuals are (a) "patternless," (b) have few or no "extreme" values, and (c) are in some sense "smaller" than those in other models. If these tests are met, then there is a strong temptation to say, "Fine, let's ship it."

A more sophisticated traditionalist may still have some concern about bias (of several flavors), but let's leave that concern for a later time. There are more than a few traditionalists who would do exactly as I have suggested.

2. But Wait — There's a Whole 'Nuther World Out There

With enough parameters or enough ad hoc fiddling, it's not that hard to meet the three criteria of the traditionalist, but the model still may be a horrifying misrepresentation of reality. Here is a suggestion by Andreas Buja (October 10,

Date: Fall 2007.
1991 *Mathematics Subject Classification.* Primary: xxxx; Secondary: xxxx.
Key words and phrases. Models.

2007) that goes an important step further. It's a clear simple suggestion that deserves to be followed every single time.

The suggestion is that one should not just look at residuals, but use instead use the model to generate "new data" y^* and see if the generated values pass the "sniff tests" when compared to y. Specifically, one fixes the covariates, generates new noise terms ϵ^* from the theoretical error model, and calculates

$$y^* = F(x, \hat{\theta}, \epsilon^*).$$

Now one falls back on the humblest of EDA tool, like the qqplot. Does the replicate y^* look anything like the original y? If not, you must admit that the model does not offer us a useful representation of reality. One should not look to such a model for guidance about things that really matter.

Side Note:

The process that Andreas suggests may have a name, but I don't know it. It's a bit like the parametric bootstrap, except here we're are not specifically looking at features of $\hat{\theta}$, such as the distribution or the bias. This is a more down-home thing. I'd rather not call it the parametric bootstrap, unless forced. To do so would collapse some distinctions that I think are useful.

Up-Date:

Shane points out that the general idea of generating new data y^* to see if the generated values pass the 'sniff tests' when compared to y is "well- established within the Bayesian community, but not necessarily well- practiced." Shane also notes that Chapter 6 of the Gelman et.al. "Red Book" focuses on model checking and sensitivity analysis, which for a Bayesian model involves generating replicate data sets from the fitted posterior predictive distribution and comparing these replicate data sets to the observed data.

3. Fitness for Use and Modeling with a Purpose

I hope to expand this note with more honest suggestions about how one should evaluate a model. I'd love to find at least one more suggestion with the potential of the one made by Andreas. Still, for the moment I will just stir in some philosophical points.

- If you are a serious user of a model (think Manhattan project) you really cannot abide the luxury of ignoring how you intend to use your model one you have it. This brings up the notion of "fitness for use" which I find to be sadly ignored in many statistical discussion.
- Some may say, "I build general methods that may be used by many people for many purposes. The issue of fitness for use is for the next fellow, not me."
- This seems somewhat reasonable, but it doesn't cover all the bases. In a lifetime of work, it's not credible at least some of ones effort should not have been done with a concrete and important purpose in mind.
- Moreover, many people have found that work with an clear and compelling purpose greatly shifts the issues that one considers to be of importance.
  (1) There is a much sharper focus on the weakest link in the application chain. If one link is almost broken, it's virtually psychotic to focus great effort on making other links stronger.

(2) There is a much sharper focus on putting all of the pieces together to get something that does some job that some well-informed serious person honestly cares about. We can debate this, but my sense is that in most honest practical situations it is EDA and model building that is most important. Most feasible methods of estimation will agree in most reasonable contexts. When the methods of estimation do not agree, what one discovers is almost always some structural feature of model inadequacy — such as a likelihood surface that no one could ever be happy about.

4. Continuing the Discussion

This is what I have for now. Mainly I wanted to put Andreas's comment on paper, so I would not forget it. It is perhaps well-known to you. That would be delightful. I wish it was used by every statistician at every possible occasion. I hope this example will motivate you to tell me what else should be done — especially what else **must be done**. I am happy to be the chronicler of this conversation.

Just imagine what it would be like if our students all probed their work with tools like the one Andreas suggests. Our students would stand out like rare gems.

Department of Statistics, Wharton School, University of Pennsylvania, Huntsman Hall 447, University of Pennsylvania, Philadelphia PA 19104

E-mail address: steele@wharton.upenn.edu