DOES THE MODEL MAKE SENSE? PART II: EXPLOITING SUFFICIENCY

J. MICHAEL STEELE

ABSTRACT. This class note is not for publication — at least anytime soon or in anything like the present form. The intention is to report on a suggestion from Andreas Buja on how one can use sufficient statistics to do model checking with EDA and diagnostic plots. Please consider contributing to the conversation.

1. The Usual Analysis — and an Unusual Question

Suppose you have some model, say a regression model for the old chestnut, the Boston Housing Data. You want to understand the price y_i of house i given a vector x_i of natural covariates. Without a thought, you trot out a linear model – not because it makes any *a priori* sense — but because that is what you know how to do. This act behind you, you ask yourself an unusual question: "Does this model make sense?"

Our friend and mentor Andreas Buja has a cool way to engage this question. Moreover, Buja's suggestion calls on a piece of theory that you probably never thought would have any practical application.

2. Sufficient Statistics — Not Just Lame Theory Anymore

For the linear model we know that $S = (\hat{\beta}, \text{SSE})$ is a complete minimal sufficient statistic. Among other things, this means that **if the model is true** then conditional on S then the n-dimensional vector $y = (y_1, y_2, ..., y_n)$ is uniformly distributed on the n-dimensional sphere with center $X\hat{\beta}$ and squared radius $r^2 = \text{SSE}$. This is elegant old fashioned linear model stuff, but it is also news you can use.

What you do now is generate new data from this conditional model. For example, let's get five new y vectors, and let's call them y^{*i} for $i = 1, 2, ..., 5$. Now, let's do the regression for each of these five new simulated data sets, and look at our favorite diagnostic plot, the scatter plot of observed y^{*i} versus the fitted values from the linear model. We get six nice plots: one for each of our five simulated data sets and one for the corresponding plot for the original data.

These plots reveal some interesting stuff. In particular, in the plot for the original data there is a goofy artifact — a straight line at two o'clock that contains a dozen or so points. This artifact comes from a truncation of one of the covariates. The interesting news is that this artifact is **not** evident in other five plots. Thus, the plain vanilla model does not acknowledge this quirk of the data.

Consequently, we see that there is at least one way to improve our earlier model. Specifically, we can look at the piece that was truncated from one of our covariates

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

as a missing variable. What we win from our diagnostic excursion is a very concrete suggestion for how we might improve our model at least a bit.

3. You Can Also Generate Tests This Way

There are uncountably many situation where we have data y and a model $\mathcal{M}$ such that one can simulate from the distribution of y given S, a sufficient statistic. This observation can be used to generate many kinds of natural tests. In particular, this recipe leads one to some natural permutation tests.

For example, consider data $v_i = (x_i, y_i)$ $i = 1, 2, ..., n$ that we model as IID vectors with unknown distribution $F(x, y)$, and suppose we wish to test the hypothesis that the two coordinates are independent; that is, we want to test that $F(x, y) = F_1(x)F_2(y)$ for some distributions F_1 and F_2. The sufficient statistics under the independent model are just the order statistics $\{x_{(i)} : i = 1, 2, ..., n\}$ and $\{y_{(i)} : i = 1, 2, ..., n\}$. We can generate "new data set" v^{*j}, $j = 1, 2, ...$, just by picking random permutations of these sequences.

To build a test, one takes some feature of merit of the data $\{v_i : i = 1, 2, ..., n\}$ say $T(v)$ the correlation (or, alternatively the rank correlation). Next, one generates a zillion new v^{*j} data sets. Naturally, the p-value is just the percentage of the values $T(v^{*j})$ that are more extreme than $T(v)$. Next case!

4. Bottom Line

In a sense, we're finding that "MCMC revolution" provides a straightforward way to view classical statistical methods and to take a step beyond them. Here the key observation is that many models that one might write down will come equipped with useful sufficient statistics — both big, like order statistics, and small like $(\hat{\beta}, \text{SSE})$. Here size does not matter. What matters is that it is often shockingly easy to simulate from the conditional distribution. This gives us lots of new shots, from informative diagnostic plots to sensible significance tests.

Some Questions:

- The significance tests from the "sufficient statistic model simulations" were a snap, but what about power? In the independence example, this is not trivial since one needs some measure of distance from the null.
- What classical data analyses might one revisit with this new point of view?
- What can be done with this paradigm in the land of hazard rate models or Kaplan-Meier models?
- How about time series models? Certainly consider the simplest cases first, but don't wait too long to consider Garch models and leverage models such as TGarch and EGarch.

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

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