Crowdsourcing data mining

Ben Klemens

24 March 2008

In in entry #005, I wrote about how dataviz tools take an extreme position in the descriptive-versus-inferential balance, by giving you the option of eyeballing every possible test, and then picking the one that works best. But you can't run *every* test, because your time is limited. The solution: crowdsource! Put your data up on iCharts[1], iFree3d[2], Many Eyes[3], Swivel[4], Timetric[5], Track-n-Graph[6], Trendrr[7], or Widgenie[8], and then let others try every test that seems sensible to them.

The dicta of the dataviz gurus offer plenty of good advice for presenting a known result. When writing a good essay, you want every sentence to help the reader understand the essay's conclusion; when producing a good plot, every inkblot should help the reader understand the plot's conclusion. In that context, having easily accessible tools where you can just drop in your data and get many types of well-designed plots is fabulous.

But the promotional copy for these sites—and even the name *Many Eyes*—suggests that these aren't just tools for effectively plotting the results of research, but allow the crowdsourcing of the search for patterns. This is where problems arise, because as per last episode, this is a recipe for finding both patterns that are and are not present.

Now, to be fair, if you really wanted to just snoop around for the best fit, the time constraint I'd alluded to above is not a particularly serious problem. It's not very hard to write a loop to try every possible combination of a set of variables and report which set provides the best fit. [I'd actually written up the try-every-regression loop as an exercise in *Modeling with Data*, but cut it because I thought it was too cynical.]

Here's my favorite interview[9] regarding this issue. The testing-oriented interviewer came as close as politely possible to asking the people who first developed the lattice display, Rick Becker and Bill Cleveland, whether this display method treads too close to data snooping:

Interviewer: OK, but there is another way to approach the study of a large

[1] `http://icharts.net`
[2] `http://ifree3d.com`
[3] `http://services.alphaworks.ibm.com/manyeyes/`
[4] `http://www.swivel.com/`
[5] `http://timetric.com/`
[6] `http://www.trackngraph.com/`
[7] `http://www.trendrr.com/`
[8] `http://widgenie.com/`
[9] `http://stat.bell-labs.com/project/trellis/interview.html`

> database: develop a statistical model and see if it fits the data. If it does fit, use the model to learn about the structure of the data.
>
> *Becker & Cleveland:* Yes, and Trellis display is a big help in doing this because it allows you to make a good guess about an initial model to fit and then to diagnose how well it fits the data. [. . .]
>
> *Interviewer:* But instead of agonizing over all those panels I could do a bunch of chi-squared tests for goodness of fit.
>
> *Becker & Cleveland:* You're joking, right? If not, we're leaving.
>
> *Interviewer:* OK, I guess I'm joking.

To a descriptive person, looking at a Trellis™ plot is nothing like looking at a matrix of goodness-of-fit statistics—that's certainly not how it feels. But the two activities are in the end closely correlated, and if a regression line looks good on the plot, it has good odds of passing any goodness-of-fit tests.

Many eyes But let me get back to the problem of crowdsourcing the process of trying every combination of variables.

Say that one researcher finds the middle ground in the descriptive/inferential range. She comes in with some idea of what the data will say, rather than waiting for the scatterplot of Delphi to reveal it, and then refines the original idea in dialog with the data (and good plots of the data). The researcher is not on a pure fishing expedition, but she is not wearing blinders to what the data has to say.

So one researcher could be reasonable—but what happens when there are thousands of reasonable researchers? When a relevant and expensive data set has been released, a large number of people will interrogate it, each with his or her own prior expectations. I've been to an annual conference attended by about a hundred people built entirely around a single data set, and who knows how many weren't able to fly out. With so many researchers looking at the same set of numbers, *every reasonable hypothesis will be tested.* Even if every person maintains the discipline of balancing data exploration against testing, we as a collective do not.

Every person was careful to not test every option, so none would seem to be mining the data for the highest statistical significance. But collectively, a thousand hypothesis tests were run, and journals are heavily inclined to publish only those that scored highly on the tests. So it's the multiple testing problem all over again, but the context is the hundreds or thousands of researchers around the planet studying the same topic. Try putting *that* into a cookbook description of a test's environment.

So our tests just aren't as powerful as we think they are, because we're not taking into account the true, collective context. Both halves of the descriptive/inferential balance are essential, but the inferential side is increasingly diluted and weakened by the scaling-up of our descriptive powers. There's no short-term solution to this one, though in an episode or two, I'll discuss some band-aids.