Statistical Modeling, Causal Inference, and Social Science

Advice on writing research articles

Posted by Andrew on 30 July 2009, 10:08 am

The American Statistical Association organizes a program in which young researchers can submit writing samples and get comments from statisticians who are more experienced writers. I agreed to participate in this program, as long as the authors were willing to have their articles and my comments posted here.

I'm going to start with my general advice after reading and commenting on the two articles sent to me. I think this advice should be of interest to nearly all the readers of this blog. Then I'll link to the articles and give some detailed comments.

General advice

Both the papers sent to me appear to have strong research results. Now that the research has been done, I'd recommend *rewriting both articles from scratch*, using the following template:

1. Start with the conclusions. Write a couple pages on what you've found and what you recommend. In writing these conclusions, you should also be writing some of the introduction, in that you'll need to give enough background so that general readers can understand what you're talking about and why they should care. But you want to start with the conclusions, because that will determine what sort of background information you'll need to give.

2. Now step back. What is the principal evidence for your conclusions? Make some graphs and pull out some key numbers that represent your research findings which back up your claims.

3. Back one more step, now. What are the methods and data you used to obtain your research findings.

4. Now go back and write the literature review and the introduction.

5. Moving forward one last time: go to your results and conclusions and give alternative explanations. Why might you be wrong? What are the limits of applicability of your findings? What future research would be appropriate to follow up on these loose ends?

6. Write the abstract. An easy way to start is to take the first sentence from each of the first five paragraphs of the article. This probably won't be quite right, but I bet it will be close to what you need.

7. Give the article to a friend, ask him or her to spend 15 minutes looking at it, then ask what they think your message was, and what evidence you have for it. Your friend should read the article as a potential consumer, not as a critic. You can find typos on your own time, but you need somebody else's eyes to get a sense of the message you're sending.

Comments on the two articles sent to me

Before looking at the articles, I have a few comments. I won't be evaluating the technical content of the work; this is a writing workshop, and I'm assuming the authors will be able to find colleagues to assess the research content. What I'll do is give the articles a quick read and give my quick thoughts. The quick read is probably the most important: I doubt most readers do more than skim, unless they have a specific need for the methods described in the article–and even then, they'll typically only know the methods are useful if they can detect that from their initial glance at the paper.

I liked both articles a lot, but I'll be giving more negative comments than positive comments, just because it's easier to pick at things that can be fixed than to explain why I like things.

1. Comparing Weighting Methods in Propensity Score Analysis, by Michael Posner and Arlene Ash.

Good, clear title. Now I'm reading the abstract . . . It has a big mistake in the second sentence, where they write:

> The propensity score method involves calculating the conditional probability (propensity) of being in the treated group (of the exposure) given a set of covariates, weighting (or sampling) the data based on these propensity scores, and then analyzing the outcome using the weighted data.

On page 4 of the article the authors explain that they view unweighted methods as special cases of weighting, with weights equal to 0 or 1. This is fine, but in the abstract, I'd add a parenthetical explaining this point. I was completely baffled on this point for four pages until I got to that explanation.

On page 2, the authors say, "Observational studies . . . have strong external validity." I'd add "can" before "have."

A couple other things: You should number your pages. It's hard to get comments from people if they can't tell you what page they are commenting on. I also recommend using section numbers.

Some silly but useful advice: go through and remove all contentless words and phrases, such as:

– "Of course"
– "Note that"
– "Interestingly"
– "very"
– "nice"
– "We can see that"
– "It is important to note that"

Give descriptive captions to all your figures and tables. For example, in Figure 1, add a sentence explaining why you call this observation "extreme."

Please, please, please don't use 1980s-style computerese such as "Distshape", "Datadist", "Nobser", Seu", "PSWIP", "Maxobs", etc. This leads to unreadable pictures such as Figure 4.

Don't forget these basic principles:

(a) Don't write something unless you expect people to read it.
(b) This principle holds for tables and figures as well. Consider Table 2. Do you want the reader to know that in line 3, Min Obs is 894? I doubt it. If so, you should make a case for this. If not, don't put it down. When an article is filled with numbers and words that you neither expect or want people to read, this distracts them from the content.

OK, now I started to flip through pretty fast. I want to get to the punchline, so I'm heading toward the end. The summary is clear–almost–but, as a reader, I resist it: the authors say they came up with two new methods that are better than all the existing approaches. Here are my quick thoughts:

– If so, this should be the focus of the paper. Rather than a bland "comparing weighting methods," the article should be entitled, "A new method for . . ." This would focus the article on these comparisons throughout.

– It's hard for me to believe that the new methods dominate the old methods. Maybe so, but I'd find the presentation more convincing if the authors gave some discussion of why the new methods work better, and–especially important–where the new methods would *not* be expected to perform well.

In summary: the article is well-written throughout, but now that I got to the end, I think they should restructure.

2. Writing sample by Leslie Odom.

It needs a title. I know, this is just a writing sample, but the title is the most important part of an article: it sets up the reader's expectations.

The abstract is descriptive, and that's good. I usually like long abstracts, but I think this one is a bit too detailed for general statistical audiences. I'd start with something more direct, for example:

> The Noel-Levitz Student Satisfaction Inventory (SSI) is a standard measure used to study college students. We study the factor structure of the SSI and find it to be essentially unidimensional. We then construct a new measure . . .

I made up the above, and I'm sure it's wrong in a bunch of places. But you get the picture: jump right in and say what you found. Also, I'd remove the last sentence from the abstract: pre-emptive apologies are not usually a good idea.

First sentence of the article: I'd rather see a method justified based on its inherent importance than based on its prevalence in journals.

OK, now I'm reading thru the literature review . . . all seems fine . . . . I'm on page 11. I think the author needs to use bold section headings and also number the sections. I

almost missed it, but on the bottom of page 11 we've shifted from a literature review to a dataset. Vivid typographical signals would help here. And now's the time to be specific. Where did the data come from??? I can't figure this out. There are lots of numbers on pages 11 and 12, but I can't figure out what's happening.

I don't want to be picky-picky about the use of the passive voice, ("courses . . . were selected for participation") but here it's leading to real confusion? *Who* selected the courses? Where were they selected from? Who are these students? I want to know.

Sometimes the hardest part of writing is simply to state what you've already done.

Here's something I don't want to go on and on about, but Table 1 has tons of information that nobody will ever need. First off, round all percentages to the nearest integre. "58.8%" is meaningless. Second, put the n=493 and 532 right in the caption, then remove all n's form the table. All that matters are the percentages. And, no, no, no, do not order the ethnicity categories alphabetically! Order, for example, in decreasing order of prevalance. (Really, I think this should all be a graph, but that's another story.)

Setting all presentational issues aside for a moment, Figures 1 and 2 should be in an appendix. They are meaningless to a general reader. And even many expert readers won't know the meaning of item 23 in content area CL, for example.

Moving on, I'm grabbed by Tables 3 and 4. Too many digits! Or, more to the point, you need to figure out what message you're trying to send, and to focus, focus, focus, focus, focus. No reader will care–or should care–that CFI for Campus Support Services is 0.952 or that RMSEA CI_90 is 0.069, or whatever. **You have to draw the trail from the scientific question, to the statistical question, to the data, to the inferences, back to the statistical and scientific questions.** A dense table of numbers doesn't do this.

I recommend reducing the number of abbreviations. SSI is central to this article, so that's ok. And PCA (for principal component analysis) is standard. But I think the article will be much more readable if you avoid abbreviations such as PA, EFA, MNE, LM, and CSSQ. These things just make me, as a reader, get frustrated as I have to flip back and forth to find all the definitions scattered about the article.

OK, moving on to the conclusion . . . There's only 1/2 page of conclusions! I'd expand this. You should expect readers to flip to the end to see what you've found and what you recommend.

The big finding seems to be that there was little evidence for multidimensional structure. I'd love to see a graph here illustrating the approximate unidimensionality. That would seem to be crucial for the finding to be believable. You can't expect the reader to trust 10 pages of chi-squared tests. What we want is graphical evidence–especially for a concept such as dimensionality which is so inherently graphical.

Summary

I thank Michael Posner and Leslie Odom for sending me their articles and being willing to share these with the general public. I hope these comments are helpful both specifically and in general.

13 Comments

1. *Bruce McCullough* says:
   July 30, 2009 at 8:39 am

   I'd add two things on the importance of the concluding section and the abstract.

   First, the abstract should mention the results. Too many abstracts expect the researcher to read the article to find out what happened. Wrong. I read the abstract to determine whether I might want to take a look at the article.

   Second, I read the conclusions first. If the abstract piques my interest, then I read the conclusions to determine whether I should invest my time in reading the article.

2. *Manoel Galdino* says:
   July 30, 2009 at 9:50 am

   Do you know of any similar initiative for young political science scholars?

   I am a PHd student and would love to read some comments on my papers from more seniors professors.
   Besides, since I am a Brazilian and my native language is portuguese, it would help me to get some tips about style.

   I already liked these comments and I think they will be useful to me.

   Thanks for sharing this.

   Manoel

3. *dcase* says:
   July 30, 2009 at 10:11 am

   Sounds like a great program. I recently completed my dissertation in economics and I found that most important advice from my advisors was not so much on the substantive side (of course, they helped there) but on the writing and framing the issue at hand. Grad students in quantitative disciplines are, by design, directed to focus on the hardcore technical stuff. However, written well, the simplest ideas may have much more impact than some of the dense stuff in places like Econometrica. We need a program like this one in economics as well.

4. *Thomas Brambor* says:
   July 30, 2009 at 11:32 am

   Thank you for sharing these tips. I am just about to rewrite a paper and will take a closer look at your suggestions beforehand.

   To Manoel: Barry Weingast from Stanford has a couple of suggestions on how to write a paper on his website:

Weingast – Caltech Rules

5. *Antonio Pedro Ramos* says:
   July 30, 2009 at 12:40 pm

   There is also some suggestions on Gary King website:

   http://gking.harvard.edu/papers/

   Maybe Professor Gelman will write his own advice some day?

6. *Weiwei Cheng* says:
   July 30, 2009 at 1:29 pm

   Nice post, handy tips. I really wish there are some similar programs in Machine Learning as well. I definitely will join.

7. *Manoel Galdino* says:
   July 30, 2009 at 3:58 pm

   Thanks everyone.
   I'll take a look at the suggested websites.
   Best Regards,
   Manoel

8. *ZBicyclist* says:
   July 31, 2009 at 5:40 am

   1. Use the simplest vocabulary possible, and consistently use the same term for the same thing.

   Example: don't refer to "the smallest order statistic" when you mean the minimum data value.

   2. Be sure you have defined every Greek and Roman letter, and every subscript, in a manner that seems overly clear to you, so it will be at least minimally clear to your readers.

   Example: Does X' mean the values of X are in order, mean the derivative of X, or mean the adjusted value of X, or the transpose of the X matrix?

9. *Brad* says:
   July 31, 2009 at 6:40 am

   I've been downloading pdf articles from Google Scholar. I've noticed that recently, perhaps in the past ten years, there has been a tendency for authors to have TWO titles. It's like this: "The validity of the cross-cultural Gini attribute: A study from sub-Saharan Africa." (I made that up). There is also a tendency to put a joke or a pun in the first or second title. Like: "The virtue of the ginned-up Gini: A study from sub-Saharan Africa."

10. *John Turri* says:
    July 31, 2009 at 9:03 am

I think the program is excellent. I would like to see the American Philosophical Association do something like it for young philosophers. It would be especially useful to those *not* in leading research departments.

One minor disagreement with this: "Some silly but useful advice: go through and remove all contentless words and phrases, such as "of course," "it is important to note," etc. such. Those phrases can sometimes help the reader know how to interpret the point being made. I agree that such phrases should be used sparingly, but if you think their occasional use promotes clarity and understanding, then I'd so keep those select few around.

11. *David* says:
July 31, 2009 at 2:02 pm

Re: the approach to the abstract. I don't work in a statistic field, but I frequently write analytical reports. It is my practice to use every sentence from the intro [maybe three paragraphs worth] as the start of a paragraph. Then I'll go back and fill in the [hopefully few] blanks to make the conclusions easier to follow. Then I'll go back and edit the copied lines for style. It's not a beauty contest, it's an attempt to communicate as clearly as I can my results and my rationale for them.

I used to be in the habit of of-coursing and now-we-see-that-ing. Now, it makes me mad when I see it.

12. *John* says:
August 2, 2009 at 10:48 am

Much thanks!
This will be my quick-info page when
writing important papers.
Page bookmarked!

13. *Fab* says:
August 2, 2009 at 11:00 pm

Simon Peyton-Jones' presentation is also helpful:

http://research.microsoft.com/en-us/um/people/sim…