Defense Feedback

April 8, 2021

Overall

- Keep this list of bullet points with all the changes I implement (I'm likely not implementing everything, so keep only those that I actually use)
- Put the added/changed sections in the diss in red font, so that Jeff can easily find them:
  `\textcolor{red}{text}`

Jeff

- hot.deck is non-parametric, which means it doesn't care about distances. Change my comments on that
- hd.ord does better, i.e. more on par with the other methods, for MNAR because the methods are all equally bad here. None of them were designed for it, so none of them do well. Make sure to include this point when I adjust the conclusions to make them more positive
- Include a little more on Type I/II errors – which one does it increase more than the other? It can't be symmetric, i.e. that both happen equally. Explore that a bit more. Jeff: It's going to be that we're underestimating Type I errors. Type II errors probably won't be affected
- Include a paragraph on the notion that different ordinal categories mean different things to different people, i.e. the subjective perception of 5-point scales isn't the same across people, and whether/how that influences measuring distances between categories
- Why did I not get stronger results in the blocking chapter? Why not the results I wanted? (All the other chapters had some form of 'why' in them, so put something here as well)

Ryan

- On p. 33 or p. 36, another example of missing data methods making a political difference is available at http://www.ryantmoore.org/files/papers/wlidd.pdf
- p. 39. For packages that do hot decking, see https://cran.r-project.org/web/views/MissingData.html. Be clear about what's done and not done as it relates to your work: "hot-deck imputation is implemented in hot.deck, FHDI and VIM (function hotdeck). StatMatch uses hot-deck imputation to impute surveys from an external dataset. impimp also uses the notion of 'donor' to impute a set of possible values, termed 'imprecise imputation'"
- Is it bad to impute values that don't exist in the data? I can see how this is trickier with categorical variables, but with continuous ones, I don't think it is, personally. E.g., imagine we have a clear, positive, low-variance linear relationship between age and income: Inc = 10,000 * Age. Respondents range from 20-50 in age. We have nonrespondents with age 55. If we believe the linearity, why not impute higher incomes for those non-respondents?

- Is there any non-response in the Lucid data? If not, is that because they were required to be completes? (In a normal (?) survey, we'd have some non-response.) Maybe just a little clarification here
- MNAR and hot decking – there is a connection here, but what is it? → use Jeff's point above that they basically all suck equally for data MNAR, rather than that hot decking is somehow much better suited to MNAR
- Equation (2.7), the one with the model for the ANES: The predictor variables here predict education. Education is on the left side here, but then becomes the right side for subsequent analysis. Are there potential problems with this in terms of bias?
- Figure 2.4, linear predictor distribution: What if this returns a distributions with big gaps between categories left and right? How would we continue with category reassignment in such a case? Would we still give the new categories sequential numerical values?

Betty

- v2: The integrals in equations (3.10) and (3.11) have new notation to me: xX^{miss} at the end of each integral I am pretty sure is supposed to be the differential dX^{miss}. Double check and update that as needed
- v2: Introduction "easily reached triple digits" → citation for that?
- v2: Figure 2.1: Explain why the sample sizes differ between rows, i.e. 14 for 2 groups, 18 for 3, 25 for 5 etc.
- v2: p. 24 "The first data set contains all observations" → Can these data sets have the same individuals?
- v2: p. 53, 54 clarify "numerous", "number of imputations", "1,000"
- v2: p. 54 clarify how interest is included in the polr treatment (confusingly worded)
- v2: p. 57 and top of p. 59 same "interest" thing
- v2: p. 68 is there anything in the default settings of hot.deck and hd.ord that might need to be tweaked to improve their performance?
- v2: p. 10 left side: Are you using the convention that Y is a random variable (R.V) and y is an observed realization of the R.V.?
- v2: p. 12 left side: How specifically are you blocking here based on v?
- v2: p.15 left side: I am trying to wrap my head around what exactly q, c, d, g, $g*$ are. For instance. is q the # of participants with covariate c taking on a value of d that was assigned to group g? What is $g*$?
- v2: p. 24 ANOVA comments on right side: Do you mean by this that you are fitting the ANOVA model using the (equivalent) regression setting?
- v3: Table 4.2: Would a summary table of counts illustrate the low #s in certain categories? What are the reference categories for each variable?
- v3: bottom p. 90 was there no actual missing data in the experiment?
- v4: On p. 12 the steps for the estimation process with blocking has step 8 repeating steps 1 to 4. Are only 1 to 4 repeated? Or should 1 to 7 be repeated?
- v4: On p. 20 equation (2.4) may have a typo when you subtract $X\beta$ in the fourth step, I think that should be $p(\mu < \eta_r - X\beta)$. (You have a $+X\beta$ but should that be minus?!) (see also comments p.20 v2)
- v4: On p. 20 equation (2.5) suddenly we have X transpose. Is that correct? And there

may be some clean-up of minus signs pending the things in equation (2.4)

- (Waiting on handwriting clarifications for a few other points)

Mike

- Conclusions are too negative, there are some cases when my method might work. There might be some substantive situations where the ordinal responses are probably mapping well from the latent variable, while in other situations that mapping may not be straightforward. Adjust diss to sound more like the presentation here
- Point conclusion in a more positive direction, suggest directions that methodologists should explore, i.e.: that there is the need for further exploration of this problem to get us to a point where there either is better theory that people can use to design survey questions for latent variables — or minimally get us to a point where we might have a good diagnostic test for when we should assume the ordinal responses might map well to a latent variable and when they might not
- Mike also mentioned General Additive Models and Bayesian shrinkage approaches, which I might include when I point to further directions to explore (if I want)