Re: Potential problem with the OP model …

From: **Jeff Gill** | jgill@american.edu Monday, May 27, 2019, 14:36

To: **Simon Heuberger** | sh6943a@american.edu

This is misguided or misinterpreted. The whole point is that the first stage of polr() uses covariates to re-bin education: the education that people should have had rather than the bin they selected. I'm guessing that the comment from Ryan was misunderstood. I'm feeling like you are rushing this simulation process to get something done before leaving and not sitting back and being more reflective. It's important to step back here and think about the big picture...

- categorical assignments are dependent on the man-made survey instrument's question, not really on the latent underlying true measure

- therefore there can be measurement error

- so let's model the outcome using a specification that makes the right underlying assumption

- then we move on to blocking with more informed variable definitions

So differences of importance emergence when the information fed into blocking differs from a more naive approach. These differences are observed in different regression tables at the end of the process.

So don't rush to finish something today. The break from your daily routine and the lack of good internet should provide exactly what I'm recommending: stepping back a little before jumping into coding 100%.

> From: **Simon Heuberger** | sh6943a@student.american.edu To: **JeffGill** Sunday, May 26, 2019, 11:51
>
> Hi Jeff,
>
> Remember how Ryan said in the meeting that I should manually assign people's original education categories to the eventual new ones determined by polr(), to make sure that no one was assigned wrongly? I took that to mean that I assign the original categories to the new ones with ifelse():
>
> df$education.test <- as.factor(ifelse(df$education == "Up to 1st" | df$education == "1st-4th" | df$education == "5th-6th" |

df$education == "7th-8th" | df$education == "9th" | df$education == "10th" |
df$education == "11th" | df$education == "12th" | df$education == "HS grad", "HS grad",
ifelse(df$education == "Some college", "Some college",
ifelse(df$education == "Associate", "Associate",
ifelse(df$education == "Bachelor's", "Bachelor's", "Master's")))))
df$education.test <- factor(df$education.test, levels = levels(df$education.new))
levels(df$education.test)
[1] "HS grad" "Some college" "Associate" "Bachelor's" "Master's"

The resulting distribution (attached) looks very different from the one resulting from polr() (also attached). I also attached the original distribution, for completion's sake. I'm now worried that the whole model doesn't work … I hope you will tell me that I did something wrong here and we're all good. Thoughts?

Thanks
Simon

From: **Jeff Gill** | jgill@american.edu Monday, May 27, 2019, 21:55

To: **Simon Heuberger** | sh6943a@american.edu

I'll have respond tomorrow in detail.

-Jeff

Sent from my iPhone and therefore brief.

From: **Simon Heuberger** | sh6943a@student.american.edu Monday, May 27, 2019, 12:11

I see your point about not rushing through things just to get something done. And it is probably a good thing that I will be off things for a bit.

I ran what I thought Ryan meant. Since what I ran is misguided, I must have misunderstood what he meant — he wouldn't suggest I do something that's not methodologically sound.

Could you help me with a substantive question on this, to help me see the big picture? People select an education bin, based on their life. So someone who went to school up to 4th grade selects 4th grade. Through polr(), he gets put in a different bin, say Associate. Are we saying that this person should have had the education bin Associate, rather than the bin 4th grade, because that's where he is supposed to be, given the underlying latent continuous variable? If we are saying that, aren't we giving that person an education level he doesn't really have?

I guess I thought that, up until now, polr() dropped categories we didn't need and merely collapsed them. As in, we don't separate categories for 1st through 11th grade, so let's put them all in HS grad. I didn't think people would be reassigned to completely different bins.

From: **Jeff Gill** | jgill@american.edu Monday, May 27, 2019, 08:36

This is misguided or misinterpreted. The whole point is that the first stage of polr() uses covariates to re-bin education: the education that people should have had rather than the bin they selected. I'm guessing that the comment from Ryan was misunderstood. I'm feeling like you are rushing this simulation process to get something done before leaving and not sitting back and being more reflective. It's important to step back here and think about the big picture...

- categorical assignments are dependent on the man-made survey instrument's question, not really on the latent underlying true measure

- therefore there can be measurement error

- so let's model the outcome using a specification that makes the right underlying assumption

- then we move on to blocking with more informed variable definitions

So differences of importance emergence when the information fed into blocking differs from a more naive approach. These differences are observed in different regression tables at the end of the process.

So don't rush to finish something today. The break from your daily routine and the lack of good internet should provide exactly what I'm recommending: stepping back a little before jumping into coding 100%.

> From: **Simon Heuberger** To: **JeffGill** Sunday, May 26, 2019, 11:51
>
> Hi Jeff,
>
> Remember how Ryan said in the meeting that I should manually assign people's original education categories to the eventual new ones determined by polr(), to make sure that no one was assigned wrongly? I took that to mean that I assign the original categories to the new ones with ifelse():
>
> df$education.test <- as.factor(ifelse(df$education == "Up to 1st" | df$education == "1st-4th" | df$education == "5th-6th" | df$education == "7th-8th" | df$education == "9th" | df$education == "10th" |

> df$education == "11th” | df$education == "12th" | df$education == "HS grad", "HS grad",
> ifelse(df$education == "Some college", "Some college",
> ifelse(df$education == "Associate", "Associate",
> ifelse(df$education == "Bachelor's", "Bachelor's", "Master's")))))
> df$education.test <- factor(df$education.test, levels = levels(df$education.new))
> levels(df$education.test)
> [1] "HS grad" "Some college" "Associate" "Bachelor's" "Master's”
>
> The resulting distribution (attached) looks very different from the one resulting from polr() (also attached). I also attached the original distribution, for completion’s sake. I’m now worried that the whole model doesn’t work … I hope you will tell me that I did something wrong here and we’re all good. Thoughts?
>
> Thanks
> Simon
