Causal Inference with Ordinal Outcomes

- Political science research routinely use ordinal outcome to study outcomes lies on latent space. Quote articles interpreting / mentioning / theorizing on latent outcome spaces.
- Common ways: 1. cardinalization 2. binarization 3. ordinal regression with distributional assumptions.

Problem setting – Latent variable framework and causal estimand

- DGP on a latent space.

$$Y_i^* = f(X_i, \beta) + \tau D_i + \varepsilon_i$$

$$Y_i = j \text{ if } \alpha_{j-1} < Y^* \leq \alpha_j$$

- Introduce reporting funciton. Common reporting function assumption -> 1. what it means 2. hint this will be relaxed later.
- Introduce potential outcome framework and ATE on a latent space.
- Introduce common causal inference assumptions: 1. SUTVA 2. Ignorability 3. Positivity.
- Show if the latent outcome(Y^*) is directly observed, ATE is identified by the assumptions.

Estimand 1: ATE $= \tau$ – Up to scale identification

- Introduce up to scale identification. (No further assumption needed). Intuitively, since we cannot observe Y^*, there is no good way to know the scale and location of Y^*. -> It can be 0 to 1, or -1 to 1.
- Mathematically show up to scale identification. Since the observed probability is invariant to positive affine transformations, the scale and location cannot be pinned down. Suppose the latent space scaled by a constant $c > 0$ and location adjusted by $a \in R$:

$$\tilde{Y}_i^* = a + cY_i^*, \qquad \tilde{\alpha}_j = a + c\alpha_j.$$

Then the two probability regarding the original latent outcome Y_i^* and the probability regarding the scaled and location adjusted latent outcome $\tilde{Y}_i^*$ is indistinguishable:

$$P(Y = j) = P(\alpha_j < Y^* \leq \alpha_j) = P(\tilde{\alpha}_j < \tilde{Y}_i^* \leq \tilde{\alpha}_j)$$

-> Theorem 1.

Theorem 1 (Scale and Location Non-Identification) *Any parameter vector* (α, β, τ) *and its positive affine transformation* $(a + c\alpha, c\beta, c\tau)$ *with* $c > 0$ *induce the same distribution for the observed ordinal outcome* Y_i *conditional on* (X_i, D_i)*. Consequently, the absolute scale and location of the latent outcome* Y_i^**, and hence of ATE***, are not identified from ordinal data alone.*

- Signs are identified, but the magnitude is only comparable inside a model. Cannot directly compare coefficients across models and across samples.

Transformation of ordinal outcomes: cardinalization and binarization

- Two of the most common approaches.
- Cardinalization imposes arbitrary numeric labels, which may distort the sign and magnitude of ATE.
- Binarization imposes arbitrary grouping, which may distort the result depending on how we group ordinal outcomes and lose a lot of statistical power (usually need 2-3 times bigger samples).
- Small numerical example
  - show that the common cardinalization (1-5) can sometimes correct and sometimes wrong, based on the reporting functions.
  - show that binarization result depend on the grouping scheme.

Normalization

- By Theorem 1, it is clear that the model can only be identified up to scale.
- Therefore the treatment effect also can be identified up to scale.
- However, we can nonetheless construct normalized version of ATE, and this quantities can still be point-identified, since the scale and location unidentifiability applies across the whole latent space. This paper discusses three common normalizations: 1) error scale normalization, 2) index scale normalization, and 3) anchoring.

Estimand 2: error scale normalized ATE – $\tau_\varepsilon = \frac{\tau}{\sigma_\varepsilon}$

- Since by Theorem 1, the treatment effect τ is identified up to scale: $c\tau$. The error (the residuals) can be identified up to scale and location: $a + c\varepsilon$. (Or succintly, $(\tau, \varepsilon) \sim (c\tau, a + c\varepsilon)$)
- Define error scale normalized ATE as $\tau_\varepsilon := \frac{\tau}{\sigma_\varepsilon}$.
- Since τ_ε is invariant to affine transformations ($\frac{c\tau}{c\sigma_\varepsilon} = \frac{\tau}{\sigma_\varepsilon}$), it can be point-identified.
- The error scale normalized ATE can be thought of as the treatment effect in units of the error standard deviation. This is also similar to Cohen' D, in the sense that the effect is measured relative to the noise

(the residuals / the error).

Estimand 3: index scale normalized ATE – $\tau_{index} = \frac{\tau}{\sigma_{m(X_i)}}$

- Let $m(X_i) = f(X_i, \beta)$ be the *index* of the model. This index can be identified up to scale by Theorem 1. Then the standard deviation of it: $\sigma_{m(X_i)}$
- Again, by normalizing the treatment effect by the standard deviation of the index will be point identified: $\tau_{index} = \frac{c\tau}{c\sigma_{m(X_i)}} = \frac{\tau}{\sigma_{m(X_i)}}$.
- This can be interpreted as the treatment effect relative to overall variability caused by other covariates X_i.

Estimand 4: Anchoring – $\tau_{anchor} = \frac{\tau_1}{\tau_2}$**.**

- Suppose there are multiple independent variable, the effect of which can which can be identified up to scale. That is, now we have multiple (randomized) binary treatment D_{ij} where $j \geq 2$.
- All treatments effects will be identified up to the common scale by Theorem 1: $c\tau_j$
- The size of each treatment effect can be expressed relative to another, and these relative effects (ratios) are point-identified.
- We can then define and point-identify the anchored treatment effect, which measures treatment effects relative to an anchor. If we let one treatment to work as an *anchor*, then we can interpret every other treatment effects relative to that effect, for example, we can anchor the treatment effect of D_2, to be 1 ($\tau_2 = 1$) and all other anchored treatment effects can be interpreted relative to this anchor.
- This strategy can be easily adopted by the conjoint experiments, interpreting effects of attributes relative to one specific attribute.

Probability

- As it is shown in the derivation of the Theorem 1, the probability for each categories ($P(Y_i = j|D_i = d)$) are point-identified . We can define an alternative causal effect with these probabilities: category specific treatment effect $= P(Y_i = j|D_i = 1) - P(Y_i = j|D_i = 0)$.
- This is free from Y^* scale and location non identifiability, since this is not defined on the latent scale.
- However, these cannot be summarized as *a single statistic*, unless assumption on the weights of each categories; for example, it is equally important to increase the probability for each categories.

Heterogeneous reporting function

- The common reporting function assumption can be relaxed by making the thresholds depending on X_i. This is a generalization of ordinal regression.
- The only assumption we need is that the treatment is independent from the thresholds. Otherwise, the treatment effect varies across categories; treatment effect between category 1 and category 2 $\neq$ that of 2 and 3. (-> Heterogenous treatment effect, and cannot be summarized as overall ATE, since we do not know the weights (distances) for each categories)

Estimation

- Intro: parametric ordinal regression: logit, probit, most IRT models

Parametric baseline models

- Distributional assumption & MLE estimation formula
- Consequences of distributional assumption -> inconsistent estimation (converge to pseudo value). Sign can be flipped for finite samples. Usually, asymptotically identify the sign.
- This motivates relaxing distributional assumption based on density estimation techniques.

Semiparametric KDE based estimation

- Klein and Sherman (2002)
- KDE is nonparametric density estimation method. -> assymptotically converges to the true density with moderate bandwiths. (rule of thumb bandwidths). Triming, adaptive bandwidth selection can be applied.

Normalizing Flow based estimation

- NF is a generative model that can map complex distribution to simple base distribution with sequence of invertible and differentiable transformation. The existence of such transformation is guaranteed theoretically (Guasianization).
- Introduce math part: change of variables formula.
- The transformation can be designed to be fully parametric, and jointly estimated with the main model parameters with MLE. -> Maximum likelihood theory guarantees asymptotic properties.

Practical suggestions

- Parametric logit probit can be inconsistent if distributional assumptions are violated.
- KDE and flow-baed estimation provides more flexible structure to consistent estimation.