Some Recent Research Ideas

Nikita Pavlov

Penn State University

September 2024

Partial Identification as a Way to Handle Absent Data in Regression Analysis

- Let (Y, X, Z) be random variables equipped with some unknown joint distribution Q on $\mathcal{Y} \times \mathcal{X} \times \mathcal{Z} \subset \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$;
- Let $E_Q(Y|X = x, Z = z) = f(x, z; \theta)$ with f being some known function depending on a parameter $\theta \in \Theta \subset \mathbb{R}^k$
- Suppose we have two datasets independently drawn from Q: $\{Y_i^A, Z_i^A\}_{i=1}^N$ and $\{X_j^B, Z_j^B\}_{j=1}^M$. What can we learn about θ?

Partial Identification as a Way to Handle Absent Data in Regression Analysis

- Common approach frequently used to handle the problem described before is to predict X_i^A by Z_i^A from information in $\{X_j^B, Z_j^B\}_{j=1}^M$. That is:
  - Assume that $E(X|Z = z) = g(z; \gamma)$ for some known g and unknown $\gamma \in \Gamma \subset \mathbb{R}^p$;
  - Estimate γ using $(X_j^B, Z_j^B)_{j=1}^M$;
  - In dataset A, use $g(Z_i^A; \hat{\gamma})$ instead of absent X_i^A to estimate θ.
- However, validity of the procedure described above relies heavily on functional forms of f and g;
- It requires some strong independence assumptions even when both f and g are linear (Ogburn et. al (2021)).

Partial Identification as a Way to Handle Absent Data in Regression Analysis

- Manski and Tamer (2002) focus on a similar problem where researcher also does not directly observe X_i but instead of auxiliary dataset has information on bounds of each X_i: $\{Y_i, X_{l,i}, X_{u,i}, Z_i\}_{i=1}^N, P(X_{l,i} \leq X_i \leq X_{u,i}) = 1$
- Using partial identification techniques, authors establish sharp bounds for θ: $\mathcal{H}[\theta] = \{\theta \in \Theta : f(X_l, Z; \theta) \leq E_Q[Y|X_l, X_u, Z] \leq f(X_u, Z; \theta) \; a.s.\}$
- Perhaps, instead of using $\{X_j^B, Z_j^B\}_{j=1}^M$ to predict X_i^A based on Z_i^A, we can use $\{X_j^B, Z_j^B\}_{j=1}^M$ to estimate bounds of $X_i^A|Z_i^A$ and then use Manski and Tamer (2002) approach to estimate sharp interval of θ.

Partial Identification as a Way to Handle Absent Data in Regression Analysis

- To illustrate the idea, suppose that instead of having $\{Y_i^A, Z_i^A\}_{i=1}^N$ and $\{X_j^B, Z_j^B\}_{j=1}^M$, we had only $\{Y_i, Z_i\}_{i=1}^N$ and knowledge of support of $X|Z = z$ for every realization of Z;
- Define the support of X as follows:

$$supp(X) = \{X \in \mathbb{R} : \forall r > 0, P_X(B(x, r)) > 0\}$$

- Assume that $supp(X)$ is bounded;
- Define $X_{l,i} = inf[supp(X|Z = Z_i)]$ and $X_{u,i} = sup[supp(X|Z = Z_i)]$. Now we may use Manski and Tamer techniques to estimate bounds of θ.

Partial Identification as a Way to Handle Absent Data in Regression Analysis

- In reality, we do not have perfect knowledge of support of $X|Z$ and only observe $\{X_j^B, Z_j^B\}_{j=1}^M$;
- Still, perhaps we may use $\{X_j^B, Z_j^B\}_{j=1}^M$ to come up with a procedure that would asymptotically resemble perfect knowledge of bounds of $X|Z$.

Idea Sketch

- Assume that random vector Z has finite support:

$$supp(Z) = \{z_1, ..., z_I; P(Z = z_i) > 0, \forall i \in \{1, ..., I\}\}$$

- Partition $\{X_j^B, Z_j^B\}_{j=1}^M$ into I sets defined as follows:

$$\forall i \in \{1, ..., I\}, A_i = \{(x, z) \in \{X_j^B, Z_j^B\}_{j=1}^M : z = z_i\}$$

- Define $\hat{X}_{l,i} = inf_x A_i$ and $\hat{X}_{u,i} = sup_x A_i$. It could be shown that $\hat{X}_{l,i} \xrightarrow{p} inf[supp(X|Z = z_i)]$ and $\hat{X}_{u,i} \xrightarrow{p} sup[supp(X|Z = z_i)]$
- Hence, Manski and Tamer (2002) results could (hopefully) be applied without much change;
- If some dimension of Z is continuous/unbounded, we may use various discretization techniques.

Difference-in-Differences when CVaR is an object of interest

- Oftentimes policymaker may be interested in estimating the effects of treatment on tails of distribution of respective treatment group (e.g., vaccine trials, change in insurance policy, etc.);
- Natural measure of weighted well-being in one of the tails of outcome distribution is CVaR (superquantile);
- Existing approach to superquantile regression (e.g., Rockafellar et. al 2014) does not allow for tractable characterization of ATT and hypothesis testing.

Difference-in-Differences when CVaR is an object of interest

- Athey and Imbens (2006) present a technique using which it is possible to estimate entire counterfactual distribution of outcomes for the treatment group;
- As a result, their approach allowed to obtain estimator of α-quantile ATT and deduce its asymptotic distribution;
- The same could be done for CVaR both for continuous and discrete outcomes.

Athey and Imbens (2006) Setting

Let $G \in \{0, 1\}$ denote group and $T \in \{0, 1\}$ denote time.

Further, let $Y^I_{g,t}$ and $Y^N_{g,t}$ be random outcomes upon receiving and not receiving treatment conditional on $G = g, T = t$, respectively.

Authors are interested in estimates of ATT and α-quantile ATT:

$$ATT = E[Y^I_{1,1}] - E[Y^N_{1,1}]$$

$$Q_\alpha(ATT) = F^{-1}_{Y^I_{1,1}}(\alpha) - F^{-1}_{Y^N_{1,1}}(\alpha)$$

Athey and Imbens (2006) Results

Athey and Imbens show that under certain assumptions we can recover counterfactual distribution of outcomes for the treatment group:

$$F_{Y^N_{1,1}}(y) = F_{Y^N_{1,0}}(F^{-1}_{Y^N_{0,0}}(F_{Y^N_{0,1}}(y)))$$

As a result:

$$ATT = E[Y^I_{1,1}] - E[Y^N_{1,1}] = E[Y^I_{1,1}] - E[F^{-1}_{Y^N_{0,1}}(F_{Y^N_{0,0}}(Y^N_{1,0}))]$$

$$Q_\alpha(ATT) = F^{-1}_{Y^I_{1,1}}(\alpha) - F^{-1}_{Y^N_{1,1}}(\alpha) = F^{-1}_{Y^I_{1,1}}(\alpha) - F^{-1}_{Y^N_{0,1}}(F_{Y^N_{0,0}}(F^{-1}_{Y^N_{1,0}}(q)))$$

Both expressions have finite-sample counterparts with tractable asymptotic characterization.

Athey and Imbens (2006) application to CVaR

Recall that for continuous Y, $CVaR_\alpha = \frac{1}{1-\alpha} E[I_{\{Y \geq F_Y^{-1}(\alpha)\}} Y]$.

Hence, given the results described above, in case when Y is continuous it is straightforward to characterize $CVaR_\alpha(ATT)$, find its finite-sample counterpart and work out its asymptotic distribution.

Then, more advanced techniques could be adapted to do the same for discrete Y.