Online Appendix to: Intergenerational Mobility in India: New Methods and Estimates Across Time, Space, and Communities (Asher, Novosad, Rafkin, 2021)

February 2021

1 Methods

This Online Appendix provides details about the analytical and numerical procedures to bound the CEF and functions of the CEF. These methods are straightforward applications of Novosad et al. (2020). In Appendix 1.1 and Appendix 1.2, we reproduce the text of several propositions contained in Novosad et al. (2020) for ease of reference, but relegate the proofs to Novosad et al. (2020). In Appendix 1.3, we explain the simple procedure to adapt the numerical techniques in Novosad et al. (2020) to this setting.

Relationship to Novosad et al. (2020). Novosad et al. (2020) is concerned with estimating bounds on $E(y|x=i)$ and various functions of that CEF, where x is an interval-censored adult education rank and y is that same adult's mortality rate. This paper is concerned with the same mathematical problem, where x is an interval-censored parent education rank and y is a measure of child socioeconomic status. Note that the monotonicity condition here is similar to that in Novosad et al. (2020). Here, we assume child status is *increasing* in parent education rank; Novosad et al. (2020) assumes adult survivalship is *increasing* in adult education rank.

1.1 Formal Statement of Proposition 1

Let the function $Y(x)=E(y|x)$ be defined on [0,100]. Form the set of non-overlapping intervals $[x_k,x_{k+1}]$ that cover [0,100] for $k\in\{1,...,K\}$. We seek to bound $E(y|x)$ when x is known to lie in the interval $[x_k,x_{k+1}]$; there are K such intervals. Suppose that

$$x\sim U(0,100), \qquad \text{(Assumption U)}$$

and define

$$r_k:=\frac{1}{x_{k+1}-x_k}\int_{x_k}^{x_{k+1}}Y(x)dx.$$

Adopt the following assumptions from Manski and Tamer (2002):

$$\text{Prob}(x \in [x_k, x_{k+1}]) = 1. \qquad \text{(Assumption I)}$$

$$E(y|x) \text{ must be weakly increasing in } x. \qquad \text{(Assumption M)}$$

$$E(y|x,\, x \text{ is interval censored}) = E(y|x). \qquad \text{(Assumption MI)}$$

Proposition 1. *Let x be in bin k. Under assumptions M, I, MI (Manski and Tamer, 2002) and U, and without additional information, the following bounds on $E(y|x)$ are sharp:*

$$\begin{cases} r_{k-1} \leq E(y|x) \leq \frac{1}{x_{k+1}-x}((x_{k+1}-x_k)r_k - (x-x_k)r_{k-1}), & x < x_k^* \\ \frac{1}{x-x_k}((x_{k+1}-x_k)r_k - (x_{k+1}-x)r_{k+1}) \leq E(y|x) \leq r_{k+1}, & x \geq x_k^* \end{cases}$$

where

$$x_k^* = \frac{x_{k+1}r_{k+1} - (x_{k+1}-x_k)r_k - x_k r_{k-1}}{r_{k+1}-r_{k-1}}.$$

1.2 Formal Statement of Analytical Bounds on μ_a^b

We now state a proposition, also contained in Novosad et al. (2020), that permits us to bound μ_a^b.

Define

$$\mu_a^b = \frac{1}{b-a}\int_a^b E(y|x)di.$$

Let Y_x^{min} and Y_x^{max} be the lower and upper bounds respectively on $E(y|x)$ given by Proposition 1. We seek to bound μ_a^b when x is only known to lie in some interval $[x_k, x_{k+1}]$.

Proposition 2. *Let $b \in [x_k, x_{k+1}]$ and $a \in [x_h, x_{h+1}]$ with $a < b$. Let assumptions M, I, MI (Manski and Tamer, 2002) and U hold. Then, if there is no additional information available, the following bounds are sharp:*

$$\begin{cases} Y_b^{min} \leq \mu_a^b \leq Y_a^{max}, & h = k \\ \frac{r_h(x_k-a)+Y_b^{min}(b-x_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(x_k-a)+r_k(b-x_k)}{b-a}, & h+1 = k \\ \frac{r_h(x_{h+1}-a)+\sum_{\lambda=h+1}^{k-1} r_\lambda(x_{\lambda+1}-x_\lambda)+Y_b^{min}(b-x_k)}{b-a} \leq \mu_a^b \leq \frac{Y_a^{max}(x_{h+1}-a)+\sum_{\lambda=h+1}^{k-1} r_\lambda(x_{\lambda+1}-x_\lambda)+r_k(b-x_k)}{b-a}, & h+1 < k. \end{cases}$$

1.3 Bounding Functions of the CEF

We now describe our procedure for bounding arbitrary functions of the CEF. We conduct the following process.

1. Consider the set of CEFs that can: (a) match the observed mean levels of child rank within each parent rank bin, and (b) are consistent with any additional assumptions (e.g., monotonocity and/or smoothness assumptions).
2. For every CEF in this set, generate a function of the CEF. Report the maximum and minimum value of this function, collecting values over all CEFs in this set.

Formally, index interval-censored bins by k: define the non-overlapping intervals $[x_k, x_{k+1}]$ that cover [0,100] for $k \in \{1,...,K\}$. Then define $\{r_k\}_{k=1}^K$ as the set of observed mean values of y over each bin $k \in \{1,...,K\}$. Further define $S(\{r_k\}_{k=1}^K)$ to be the collection of CEFs that is consistent with these bin means and any desired auxiliary assumptions. For example, noting that x is uniformly distributed, we can put:

$$S\left(\{r_k\}_{k=1}^K\right) = \left\{Y(x) \middle| \; Y(x) \text{ is weakly increasing}\right\} \bigcap \left\{Y(x) \middle| \; \frac{1}{x_{k+1}-x_k}\int_{x_k}^{x_{k+1}} (Y(x) - r_k(x))dx = 0, \text{ for all } k\right\}. \tag{1.1}$$

Our objective is to bound $\gamma = \gamma(Y)$, some function of the CEF. In particular, we face the following constrained optimization problem to obtain the maximum and minimum values of γ:

$$\gamma^{\min} = \min_{Y \in S\left(\{r_k\}_{k=1}^K\right)} \tilde{\gamma}(Y) \tag{1.2}$$

$$\gamma^{\max} = \max_{Y \in S\left(\{r_k\}_{k=1}^K\right)} \tilde{\gamma}(Y). \tag{1.3}$$

Novosad et al. (2020) provide details on the numerical techniques used to solve this problem. The bounds we report are the set $[\gamma^{\min}, \gamma^{\max}]$. We now describe how we apply this process in the case of the rank-rank gradient and with curvature constraints.

Rank-rank gradient. In the case of the rank-rank gradient, we let γ represent the slope of the linear approximation to the CEF. That is, fixing a CEF $Y(x)$, define

$$(\gamma, b) := \operatorname*{argmin}_{\gamma', b' \in \mathbb{R}} \int_0^{100} (Y(x) - \gamma' x + b')^2 dx.$$

Curvature constraints. In the case of reporting the CEF with curvature constraints, we simply define $p_x(Y)$ to be the value of the CEF at a given x. We define S to be the set of CEFs that are consistent with monotonicity and a second derivative that lies below a given magnitude in absolute value. In the case where there are no CEFs that precisely match the bin means (e.g., for a small enough curvature constraint), we solve a modified problem described formally in Novosad et al. (2020). Define $T\big(\{r_k\}_{k=1}^K\big)$ to be the set of CEFs that (a) minimize some distance metric between the bin means and the CEF, and (b) are consistent with the observed bin means and extra assumptions (monotonicity and curvature). In particular, for distance metric $\|\cdot\|$, define

$$M(Y) \coloneqq \int_0^{100} \|Y(x) - r_k(x)\| dx,$$

where $r_k(x) \coloneqq r_k$ if $x \in [x_k, x_{k+1}]$. $M(Y)$ is the weighted distance between a given CEF Y and the bin means $\{r_k\}$. Then define

$$T\big(\{r_k\}_{k=1}^K\big) = P \bigcap \left\{ Y(x) \middle| \left(\int_0^{100} \|Y(x) - r_k(x)\| dx \right) \leq \underline{M(P)} \right\}, \tag{1.4}$$

for

$$\underline{M(P)} \coloneqq \min_{Y \in P} M(Y)$$

and

$$P \coloneqq \Big\{ Y(x) |\ Y(x) \text{ is weakly increasing and has second derivative less than } \underline{\mathrm{C}} \Big\}.$$

Put otherwise, we find the minimum distance between the CEFs and observed bin means, as long as these CEFs obey certain properties. Then, we find the set of CEFs that obey these properties and satisfy this minimum distance. If there are CEFs that precisely meet the bin means, then $T = S$.

Finally, we report:

$$p_x^{\min} = \min_{Y \in T\left(\{r_k\}_{k=1}^K\right)} \tilde{p}_x(Y) \tag{1.5}$$

$$p_x^{\max} = \max_{Y \in T\left(\{r_k\}_{k=1}^K\right)} \tilde{p}_x(Y). \tag{1.6}$$

In practice, we choose the mean-squared error as the distance metric.

References

Manski, Charles F. and Elie Tamer, “Inference on Regressions with Interval Data on a Regressor or Outcome,” *Econometrica*, 2002, *70* (2), 519–546.

Novosad, Paul, Charlie Rafkin, and Sam Asher, “Mortality Change Among Less Educated Americans,” 2020. Working Paper.