Monotone Instrumental Variables

Thomas Wiemann
University of Chicago

TA Discussion # 5

November 3, 2021

Outline

- Empirical Example & Motivation
- Identification of Potential Outcome Distributions
  - Random Treatment Selection
  - Outcomes Missing-at-Random
  - Statistical Independence of Outcomes and Instruments
  - Linearity Restrictions
- Monotone Instrumental Variables
  - Monotone Treatment Selection
  - Monotone Treatment Response
- Application to Angrist and Krueger (1991)

Empirical Example & Motivation

Consider the question on the returns to schooling (e.g., Angrist and Krueger, 1991). Let $Y \in \mathcal{Y}$ and $D \in \mathcal{D}$ denote the the realized (log-weekly) wages and years of schooling, respectively. Let $Y(d)$ denote the potential (log-weekly) wages under d years of schooling. A key object of interest is the distribution of potential outcomes $P(Y(d))$ for each $d \in \mathcal{D}$. (For ease of analysis, assume here and throughout that $P(D = d) > 0, \forall d \in \mathcal{D}$.)

Question: What information on $P(Y(d))$ can we gain from the observed joint distribution $P(Y, D)$?

We have

$$\begin{aligned} P(Y(d)) &= P(Y(d)|D = d)P(D = d) + P(Y(d)|D \neq d)P(D \neq d) \\ &= P(Y|D = d)P(D = d) + P(Y(d)|D \neq d)P(D \neq d). \end{aligned} \tag{1}$$

Hence the identification region of $P(Y(d))$ from the empirical data is

$$H\{P(Y(d))\} = \{P(Y|D = d)P(D = d) + \gamma P(D \neq d), \, \gamma \in \Gamma_{\mathcal{Y}}\}, \tag{2}$$

where $\Gamma_{\mathcal{Y}}$ denotes the set off all probability measures over $\mathcal{Y}$.

Empirical Example & Motivation (Contd.)

The identification region of all potential outcome distributions $\{P(Y(d)), \forall d \in \mathcal{D}\}$ given the empirical data is

$$H\{P(Y(d)), \forall d \in \mathcal{D}\} = \times_{d \in \mathcal{D}} H\{P(Y(d))\}, \tag{3}$$

where $\times$ denotes the Cartesian product.

These identification regions are sharp because the sampling process does not reveal any information on $P(Y(d)|D \neq d)$.

So how useful is $H\{P(Y(d)), \forall d \in \mathcal{D}\}$? Not very. Note that the empirical evidence alone cannot refute key hypotheses. E.g.,

- $P(Y(d)) = P(Y(d')), \quad \forall d, d' \in \mathcal{D},$
- $Y_i(d) = Y_i(d'), \quad \forall d, d' \in \mathcal{D}.$

We thus *need* to make assumptions to arrive at useful conclusions about the returns to schooling.

Empirical Example & Motivation (Contd.)

Which assumptions to maintain in their analysis is a key consideration that researchers face. So how do they decide?

Manski (2003) (p. 1) characterizes a key trade-off (or: dilemma) that researchers face when considering which assumptions to maintain:

Law of Decreasing Credibility: *The credibility of inference decreases with the strength of the assumptions maintained.*

The principle highlights that stronger assumptions yield stronger but less credible conclusions.

Today's discussion – which will explore assumptions covered Chapter 2, 7, and 9 of Manski (2003) – is motivated by this trade-off.

(Note: Manski (2003) is an excellent read. Just be careful with the notation and terminology – it may not be what you're be used to! For a lighter read, also see Manski (2013).)

Identification of Potential Outcome Distributions

You're likely (very) familiar with some possible assumptions that allow for point identification of $P(Y(d))$. One such assumption is random selection into treatments.

<u>Assumption Random Treatment Selection (RTS):</u>

$$Y(d) \perp\!\!\!\perp D, \forall d \in \mathcal{D} \tag{4}$$

Under RTS, we have

$$P(Y(d)) = P(Y(d)|D = d) = P(Y|D = d), \tag{5}$$

where the last term is observed in the data such that $P(Y(d))$ is point identified.

RTS is typically considered to be a credible assumption in randomized control trials. Arguments for its credibility in other settings require more creativity.

Identification of Potential Outcome Distributions (Contd.)

Suppose now there exists a random variable Z such that $P(D = d, Z = z) > 0, \forall (d, z) \in \mathcal{D} \times \mathcal{Z}$. Z may or may not be distinct from D.

Manski (2003) introduces such a variable Z as instrumental variable. Note however that in some contexts, it is analogous to what you may call a control variable:

<u>Assumption Outcomes Missing-at-Random (MAR):</u>

$$P(Y(d)|Z) = P(Y(d)|Z, D = d) = P(Y|Z, D = d), \ \forall d \in \mathcal{D} \tag{6}$$

MAR point identifies $P(Y(d))$. We have

$$\begin{aligned} P(Y(d)) &= \sum_{z \in \mathcal{Z}} P(Y(d)|Z = z)P(Z = z) \\ &= \sum_{z \in \mathcal{Z}} P(Y|Z = z, D = d)P(Z = z). \end{aligned} \tag{7}$$

Identification of Potential Outcome Distributions (Contd.)

A much weaker assumption is the following:

Assumption Statistical Independence of Outcomes and Instruments (SI):

$$P(Y(d)|Z) = P(Y(d)),\ \forall d \in \mathcal{D} \tag{8}$$

SI set-identifies $P(Y(d))$. The identification region is given by

$$\begin{aligned} H_{SI}\{P(Y(d))\} = \bigcap_{z \in \mathcal{Z}} \{ & P(Y|Z=z, D=d)P(D=d|Z=z) \\ & + \gamma_z P(D \neq d|Z=z),\ \gamma_z \in \Gamma_{\mathcal{Y}_z}\}, \end{aligned} \tag{9}$$

where $\Gamma_{\mathcal{Y}_z}$ denotes the set of all probability measures over $\mathcal{Y}_z$. (This allows for the possibility that $\mathcal{Y}_z \subset \mathcal{Y}$.)

Note that assumption SI is falsifiable: If SI holds, then $H_{SI}\{P(Y(d))\}$ is necessarily non-empty. Hence, the assumption cannot hold if it is empty.

Of course, non-emptiness of $H_{SI}\{P(Y(d))\}$ does not imply that SI holds.

Identification of Potential Outcome Distributions (Contd.)

Sketched derivation of $H_{SI}\{P(Y(d))\}$:

Using the empirical evidence alone, the identification region is

$$H\{P(Y(d)|Z=z)\} = \{P(Y|Z=z, D=d)P(D=d|Z=z) \\ + \gamma_z P(D \neq d|Z=z),\ \gamma_z \in \Gamma_{\mathcal{Y}_z}\}. \tag{10}$$

The identification region for the set $\{P(Y(d)|Z=z), \forall z \in \mathcal{Z}\}$ is given by the Cartesian product $\times_{z\in\mathcal{Z}} H\{P(Y(d)|Z=z)\}$.

SI states that $P(Y(d)|Z=z) = P(Y(d)|Z=z'), \forall z, z' \in \mathcal{Z}$. Hence P(Y(d)) must lie in $\bigcap_{z\in\mathcal{Z}} H\{P(Y(d)|Z=z)\}$. Any distribution within this intersection is feasible. Hence $H_{SI}\{P(Y(d))\}$ is the identification region.

Identification of Potential Outcome Distributions (Contd.)

As you may have guessed, there exists a similar assumption on mean independence of outcomes and instruments (MI).

Assumption Mean Independence of Outcomes and Instruments (MI):

$$E[Y(d)|Z] = E[Y(d)],\ \forall d \in \mathcal{D} \tag{11}$$

MI set-identifies $E[Y(d)]$. The identification region is given by

$$\begin{aligned} H_{MI}\{E[Y(d)]\} = \bigcap_{z \in \mathcal{Z}} \{ & E[Y|Z=z, D=d]P(D=d|Z=z) \\ & + y_z P(D \neq d|Z=z),\ y_z \in \mathcal{Y}_z\}. \end{aligned} \tag{12}$$

See Manski (2003) Section 7.4 for a simplified version.

Identification of Potential Outcome Distributions (Contd.)

An potentially interesting application of MI is a setting where two additional assumptions are made: a linearity assumption and a relevance condition. In particular, consider a setting where

$$Y_i(d) = \tau d + U_i, \qquad \forall d \in \mathcal{D}, \tag{13}$$

with τ unknown and fixed. Suppose further that

$$\begin{aligned} E[U|Z] &= E[U], \\ E[D|Z = z_1] &\neq E[D|Z = z_2], \quad \forall z_1 \neq z_2 \in \mathcal{Z}. \end{aligned} \tag{14}$$

It's straightforward to show

$$\tau = \frac{E[Y|Z = z_1] - E[Y|Z = z_2]}{E[D|Z = z_1] - E[D|Z = z_2]}, \tag{15}$$

which is the familiar Wald estimate. Note that knowledge of τ and the observed joint distribution $P(Y, D)$ identifies $P(U)$, and thus $P(Y(d))$.

But linearity and additive separability may not be the most credible assumptions in economic applications (hence the previous TA sessions!).

Monotone Instrumental Variables

Assumptions RTS, MAR, as well as the combined assumptions of MI with a linearity and relevance condition allow for point identification of $P(Y(d))$ (and thus $E[Y(d)]$), but are likely considered strong in many economic applications. Manski (2003) writes: "[...] *their credibility in non-experimental settings often is a matter of considerable disagreement, with empirical researchers frequently debating whether some covariate is or is not a 'valid instrument.'*" (p. 141).

Assumptions SI (and MI) *may* be credible more frequently, but the identification region $H_{SI}\{P[Y(d)]\}$ (and $H_{MI}\{E[Y(d)]\}$) may not be sufficiently tight for practically minded empiricists. Note also that in contexts with non-bounded outcomes, the identification regions do not improve on the trivial bounds.

Remarkably, Manski and Pepper (2000) introduce a set of assumptions that result in non-trivial bounds for $E[Y(d)]$ in settings with unbounded outcomes while maintaining assumptions that appear to be very widely applicable across economic contexts. (See Chapter 9 of Manski (2003) for a more general textbook treatment of this topic.)

Monotone Instrumental Variables (Contd.)

For the remainder of the discussion, assume that $\mathcal{D}$ is an ordered set.

Assumption Monotone Treatment Selection (MTS):

$$d_2 \geq d_1 \;\Rightarrow\; E[Y(d)|D = d_2] \geq E[Y(d)|D = d_1], \; \forall d_2, d_1 \in \mathcal{D} \tag{16}$$

It's worth contemplating the MTS assumption: What does it imply, for example, in the returns to schooling example?

MTS asserts that persons who choose more schooling have weakly higher wage functions than do those who select less schooling.

That's consistent with much of economic theory on education choice and wage determination. In this setting, MTS thus seems to be a fairly credible assumption.

Note that MTS is simply the MI assumption where $Z = D$ and the equality is replaced by weak inequalities.

Monotone Instrumental Variables (Contd.)

MTS partially identifies $E[Y(d)]$. The identification region is given by

$$\begin{aligned} H_{MTS}\{E[Y(d)]\} = [&E[Y|D=d]P(D \geq d) + \underline{y}P(D < d), \\ &E[Y|D=d]P(D \leq d) + \overline{y}P(D > d)]. \end{aligned} \tag{17}$$

Sketch of the derivation:

1. The data alone identifies

$$\begin{aligned} H\{E[Y(d)]\} = [&E[Y|D=d]P(D = d) + \underline{y}P(D \neq d), \\ &E[Y|D=d]P(D = d) + \overline{y}P(D \neq d)]. \end{aligned} \tag{18}$$

2. Application of MTS results in $H_{MTS}\{E[Y(d)]\}$.

Monotone Instrumental Variables (Contd.)

How does MTS help restrict the identification region of $E[Y(d)]$?

Using just the empirical data, the width of the identification region is

$$\|H\{E[Y(d)]\}\| = (\overline{y} - \underline{y})P(D \neq d). \tag{19}$$

Using the MTS assumption, we have

$$\begin{aligned} \|H_{MTS}\{E[Y(d)]\}\| = (&E[Y(d)|D = d] - \underline{y})P(D < d) \\ &+ (\overline{y} - E[Y(d)|D = d])P(D > d). \end{aligned} \tag{20}$$

MTS can thus considerably restrict the identification region. For example, let $P(D < d) = P(D > d)$, then

$$2\|H_{MTS}\{E[Y(d)]\}\| = \|H\{E[Y(d)]\}\|. \tag{21}$$

Monotone Instrumental Variables (Contd.)

Note that whenever $\underline{y}$ and $\overline{y}$ are not finite, $H_{MTS}\{E[Y(d)]\}$ does not improve upon the trivial bounds. Combining MTS with a second assumption addresses this issue.

<u>Assumption Monotone Treatment Response (MTR):</u>

$$d_2 \geq d_1 \;\Rightarrow\; Y_i(d_2) \geq Y_i(d_1),\; \forall d_2, d_1 \in \mathcal{D},\; \forall i \in \mathcal{I}, \tag{22}$$

where i denotes an individual in the population $\mathcal{I}$.

MTR and MTS, albeit expressing similar notions, are distinct assumptions. In principle, both could hold, or either, or neither.

To see this, note that MTR does *not* address the selection of high-ability individuals into higher education. MTR views education as a production process, where higher schooling maps to higher wages (or intermediary: to higher labor productivity, which results in higher wages).

Monotone Instrumental Variables (Contd.)

Remarkably, the combination of the MTS and MTR assumptions holds considerable identifying power for $E[Y(d)]$.

Let MTS and MTR both hold. The identification region of $E[Y(d)]$ is

$$H_{M\&M}\{E[Y(d)]\} = \Big[\sum_{v<d} E[Y|D=v]P(D=v) + E[Y|D=d]P(D \geq d), \sum_{v>d} E[Y|D=v]P(D=v) + E[Y|D=d]P(D \leq d)\Big]. \tag{23}$$

Importantly, $H_{M\&M}\{E[Y(d)]\}$ may be non-trivial in scenarios when the bounds $\underline{y}$ and $\overline{y}$ are not finite.

Proving this result requires some intermediate results using just the MTR assumption. We'll skip it here in the interest of time. See Manski (2003) p.147 for the details.

Monotone Instrumental Variables (Contd.)

The identification region $H_{M\&M}\{E[Y(d)]\}$ allows for the construction of bounds for a key object of interest: $E[Y(t) - Y(s)]$ – i.e., the average causal effect of a (hypothetical) shift in treatment from s to t (w/ $t > s$).

For the upper bound on $E[Y(t) - Y(s)]$, we subtract the lower bound of $E[Y(s)]$ from the upper bound of $E[Y(t)]$. In particular,

$$E[Y(t) - Y(s)] \leq \sum_{v>t} E[Y|D = v]P(D = v) + E[Y|D = t]P(D \leq t) \\ - \sum_{v<s} E[Y|D = v]P(D = v) - E[Y|D = s]P(D \geq s).$$

The lower bound can be constructed analogously.

Are these bounds sharp?

- It is jointly feasible for $E[Y(t)]$ to take its maximum and $E[Y(s)]$ to take it's minimum, thus the upper bound is sharp.
- The lower bound is always weakly negative. But notice that MTR implies $E[Y(t)] \geq E[Y(s)]$ so that the lower bound is not sharp in general.

Monotone Instrumental Variables (Contd.)

The MTS-MTR bounds can readily be implemented for discrete D. See below for a quick implementation of the (sharp) upper bound in Julia.

MTS-MTR Upper-Bound Implementation in Julia (Pt 1.)

```julia
struct myIVBound
    px # P(X=x) for all x
    Eyx # E[Y|X=x] for all x
    unq_x # An orderd set of the unique values of X

    function myIVBound(y, x)
        # Data parameters
        n = length(y)
        unq_x = sort(unique(x))
        # Calculate the empirical pmf and conditional expectation
        px = zeros(length(unq_x))
        Eyx = zeros(length(unq_x))
        for v in 1:length(unq_x)
            indx = x .== unq_x[v]
            px[v] = sum(indx) / n
            Eyx[v] = mean(y[indx])
        end
        # Return the myIVBound object
        new(px, Eyx, unq_x)
    end #MYIVBOUND
end #MYIVBOUND
```

Monotone Instrumental Variables (Contd.)

MTS-MTR Upper-Bound Implementation in Julia (Pt 2.)

```
function coef(fit::myIVBound, s, t)
    # Determine indices above and below t, s
    from_t = fit.unq_x .> t
    until_s = fit.unq_x .< s
    # Calculate the ub of E[Y(t)] and the lb of E[Y(s)]
    Eyx_px = fit.Eyx .* fit.px
    ub = sum(Eyx_px[from_t]) + (fit.Eyx[fit.unq_x .== t] *
        sum(fit.px[.!from_t]))[1]
    lb = sum(Eyx_px[until_s]) + (fit.Eyx[fit.unq_x .== s] *
        sum(fit.px[.!until_s]))[1]
    # Subtract and return the upper and lower bounds
    return ub - lb
end
```

(The implementations here were quickly coded up and are meant to be roughly within the Julia design patterns discussed in TA discussion # 1. You can probably do much better by now!)

Application to Angrist and Krueger (1991)

Let's see whether we are able to gain new insights from the MTS-MTR bounds in the returns to schooling context. This is inspired by the application in Manski and Pepper (2000), who construct upper bounds on the returns to schooling for a subsample of the National Longitudinal Survey of Youth.

Here, we consider again the sample of American men born between 1930 and 1939 considered in Angrist and Krueger (1991) that we discussed in TA discussion # 3.

We will construct upper bounds on $E[Y(t) - Y(t-4)]$ for $t \in \{4, 8, 12, 16, 20\}$ to bound the returns to different school stages in the US education system (from elementary school to post-grad degrees). (Download the reproduction code here: Econ_31720_discussion_5.ipynb.)

Note that the lower bounds are always 0 and so we can't make inferences on whether schooling has any effect. Instead, the analysis is motivated through a comparison of the upper bounds with coefficient on the returns to schooling reported in the literature.

Application to Angrist and Krueger (1991) (Contd.)

Table 1 presents bounds constructed using the sample of $N = 329,509$ American men born between 1930 and 1939 (as considered in Angrist and Krueger (1991)). The outcome is the log-weekly wage, and the endogenous variable of interest is the years of completed schooling.

Table 1: MTS-MTR Bounds for the Angrist and Krueger (1991) Data

s	t	Upper-Bound
0	4	0.879
4	8	0.647
8	12	0.395
12	16	0.448
16	20	0.415

The results imply that the average returns per year of college are not above 0.112 ($= 0.448/4$). Manski and Pepper (2000) report an upper-bound per year of college of 0.099 (using different data, of course).

Discussion

There are a number of interesting empirical applications of Manski and Pepper (2000) in the literature.

Examples in labor economics include the following:

- De Haan (2011) analyzes the relationship between parents' and childrens' schooling.
- De Haan (2017) analyzes whether financial support of secondary schools with low-ability pupils has an affect on student outcomes.

Examples in health economics include the following:

- Gundersen and Kreider (2009) analyzes the effects of food insecurity on children's' health outcomes. In a related analysis, Gundersen et al. (2012) assess the effect of lunch subsidies on childrens' health.
- Kreider et al. (2015) analyzes the relationship between dental insurance and dental care usage.

Other examples exist! (And they are much easier to find than is the case for some of the topics we discussed in previous TA sessions.)

References

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

De Haan, M. (2011). The effect of parents' schooling on child's schooling: a nonparametric bounds analysis. *Journal of Labor Economics*, 29(4):859–892.

De Haan, M. (2017). The effect of additional funds for low-ability pupils: A non-parametric bounds analysis. *The Economic Journal*, 127(599):177–198.

Gundersen, C. and Kreider, B. (2009). Bounding the effects of food insecurity on children's health outcomes. *Journal of Health Economics*, 28(5):971–983.

Gundersen, C., Kreider, B., and Pepper, J. (2012). The impact of the national school lunch program on child health: A nonparametric bounds analysis. *Journal of Econometrics*, 166(1):79–91.

Kreider, B., Manski, R. J., Moeller, J., and Pepper, J. (2015). The effect of dental insurance on the use of dental care for older adults: a partial identification analysis. *Health economics*, 24(7):840–858.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2013). *Public policy in an uncertain world*. Harvard University Press.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010.