Partial Identification

Paul Goldsmith-Pinkham

May 4, 2023

Seeking Identification

- For our questions, we need to describe an *estimand* of interest
    - This may be a causal object, or purely a statistical one
- Remember that we need to know whether this object is knowable given the data generating process
    - This is not a question of any given sample, but rather a question of our underlying assumptions and the data generating process
- Two examples:
    - When were only observe a subsample of outcomes for individuals, we assume the data is missing (completely) at random to identify the full population's average
    - To identify the average treatment effect, we assumed strict ignorability (and a few other things) to identify the effect of the treatment
- Can we assume less?

Start with a simple example

- The housing market is extremely hot and you want to know what the probability that a house sells $(Y_i = 1)$ vs doesn't sell $(Y_i = 0)$
  - In particular, you want to know if houses with backyards $(X_i = 1)$ are selling more quickly than those without $(X_i = 0)$.
- However, we only observe a sale for the set of homes that chose to go on the market $Z_i = 1$ and do not observe Y_i for $Z_i = 0$
  - Can we know the probability of sale $E(Y_i|X_i)$ for all houses?
- Formally, assume that we will have a dataset of n independent triplets (Y_i, X_i, Z_i)
  - $E(Y_i|X_i, Z_i = 1)$ is identified but $E(Y_i|X_i)$ is not without more assumptions
- But $E(Y_i|X_i)$ is *partially* identified

Simple parital identification example

- Consider the law of total probability:

$$E(Y_i|X_i) = Pr(Y_i = 1|X_i) = Pr(Y_i = 1|X_i, Z_i = 1) \times Pr(Z_i = 1|X_i) \\ + Pr(Y_i = 1|X_i, Z_i = 0) \times Pr(Z_i = 0|X_i)$$

- The problem is we cannot observe $Pr(Y_i = 1|X_i, Z_i = 0)$, by definition
  - But, we know that it cannot be greater than one, nor less than zero

- As a result, it must be that

$$E(Y_i|X_i) \in [Pr(Y_i = 1|X_i, Z_i = 1) \times Pr(Z_i = 1|X_i), \\ Pr(Y_i = 1|X_i, Z_i = 1) \times Pr(Z_i = 1|X_i) + Pr(Z_i = 0|X_i)]$$

- This is intuitive! For all the houses we don't see go for sale, at the extreme, they could either all *not* sell (the lower bound) or all sell (the upper bound).

Simple partial identification example

- This is quite powerful – we've made no assumptions on the correlation between the choice to sell and sale probability.
  - The width of this interval is important – it speaks to the informativeness of the bound (and in this case is equal to $Pr(Z_i = 0|X_i)$
  - The more likely properties are to sell, the tighter these bounds become (since the missing piece gets smaller)

- What does a missing-at-random assumption imply?
  - $Pr(Y_i = 1|X_i, Z_i) = Pr(Y_i = 1|X_i)$, which means that you can ignore the missing data and the set interval becomes a single point
  - Hence, point identification!

Simple partial identification example

- A few things worth noting from this simple example

- First, what if Y_i wasn't binary, but was real valued? E.g. sale price?
  - If the value of Y_i is unbounded, the upper bound of the set would be infinity

- However, even with unbounded Y_i, you can always bound the CDF of Y_i at different points: $F(Y_i \leq t)$ is always between 0 and 1
  - This means we can set identify quantiles of the distribution!
  - Recall that this came up when discussing censored data

- So long as there is sufficient observed data, Manski (1994) shows that the α quantile is bounded by:
  - Below: $[\alpha - P(Z_i = 0|X)]/P(Z_i = 1|X)$-quantile of the observed distribution $Pr(Y_i|X_i, Z_i = 1)$ (if $Pr(Z_i = 0|X_i) \leq \alpha$, and the min of Y_i otherwise)
  - Above: $\alpha/P(Z_i = 1|X)$-quantile of the observed distribution $Pr(Y_i|X_i, Z_i = 1)$ (if $Pr(Z_i = 0|X_i) \leq 1 - \alpha$, and the max of Y_i otherwise)

Back to generality: partial identification

- Bounds are extremely powerful – we can make substantially fewer assumptions and still potentially learn quite a bit
- However, it's worth noting that these are quite limited in practice
    - Today, I will walk through two examples where I think they are quite valuable
- Then we'll discuss why these might have had limited adoption
- We will ignore inference, although it is quite important (and has a big econometrics field associated with it)
    - See Imbens and Manski (2004) and Chernozhukov, Hong and Tamer (2007) for initial primers

Selection into employment

- Consider the problem of estimating the effect of a treatment (job training) on *wages*

- Remember that the decision to work is a sample selection problem, even when you have an RCT!

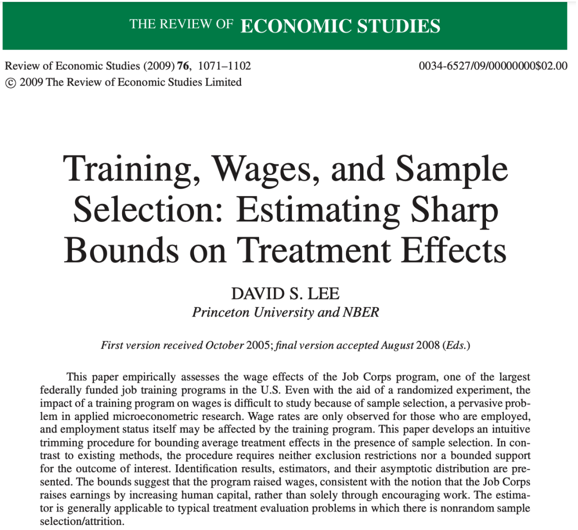
- E.g., we only observe the wages of those who choose to work – there may be endogeneous decisions to not work due to the treatment

- As a result, we have both the potential wage outcomes due to D: $(Y^*(0), Y^*(1))$ but *also* the decision to work $S(0), S(1)$. Note that we don't observe $Y^* = D_i Y^*(1) + (1 - D_i) Y^*(0)$ if $S = 0$.

- This is a serious issue that plagues almost all research designs
  - Many papers get around this by looking at "total earnings" to avoid wages

Lee Bounds (Lee 2009)

- Lee 2009 considers this problem, and proposes sharp bounds
- The problem boils down to the following issue: when you treat someone, do you change their employment status?
  - If no, the $S_i(1) = S_i(0)$ and there is no selection problem
  - If yes, the $S_i(1) \neq S_i(0)$ and the question is where do the marginal "shifters" come from and end up in the outcome distribution?
- E.g. if those who are induced to not work because they go to school b/c they get high returns from schooling, then this is negatively selecting away from the top part of the $Y^*(1)$ distribution



THE REVIEW OF ECONOMIC STUDIES

Review of Economic Studies (2009) **76**, 1071–1102 0034-6527/09/00000000$02.00
© 2009 The Review of Economic Studies Limited

Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects

DAVID S. LEE
Princeton University and NBER

First version received October 2005; *final version accepted August* 2008 (*Eds.*)

This paper empirically assesses the wage effects of the Job Corps program, one of the largest federally funded job training programs in the U.S. Even with the aid of a randomized experiment, the impact of a training program on wages is difficult to study because of sample selection, a pervasive problem in applied microeconometric research. Wage rates are only observed for those who are employed, and employment status itself may be affected by the training program. This paper develops an intuitive trimming procedure for bounding average treatment effects in the presence of sample selection. In contrast to existing methods, the procedure requires neither exclusion restrictions nor a bounded support for the outcome of interest. Identification results, estimators, and their asymptotic distribution are presented. The bounds suggest that the program raised wages, consistent with the notion that the Job Corps raises earnings by increasing human capital, rather than solely through encouraging work. The estimator is generally applicable to typical treatment evaluation problems in which there is nonrandom sample selection/attrition.

Lee Bounds (Lee 2009)

- Hopefully the problem sounds familiar (IV + LATE)
  - The solution is similar too

- Under monotonicity of selection ($S_i(1) \geq S_i(0)$, it is possible to provide sharper bounds on the treatment effect

- Let $\tau = E(Y^*(1) - Y^*(0)|S(0) = 1, S(1) = 1)$, our estimand of interest, and $p_0 = \frac{Pr(S=1|D=1) - Pr(S=1|D=0)}{Pr(S=1|D=1)}$

- Then, we can bound $\tau \in [\Delta_0^{LB}, \Delta_0^{UB}]$
  - $\Delta_0^{LB} = E(Y|D = 1, S = 1, Y \leq y_{1-p_0}) - E(Y|D = 0, S = 1)$
  - $\Delta_0^{UB} = E(Y|D = 1, S = 1, Y \geq y_{p_0}) - E(Y|D = 0, S = 1)$
  - where y_{p_0} is the p_0th quantile

Lee Bounds

- We can bound $\tau \in [\Delta_0^{LB}, \Delta_0^{UB}]$
  - $\Delta_0^{LB} = E(Y|D=1, S=1, Y \leq y_{1-p_0}) - E(Y|D=0, S=1)$
  - $\Delta_0^{UB} = E(Y|D=1, S=1, Y \geq y_{p_0}) - E(Y|D=0, S=1)$
  - where y_{p_0} is the p_0th quantile

- This has a number of cool properties:
  1. If $p_0 \rightarrow 0$, under monotonicity that means no sample selection. It also provides a test of mononicity: if covariates shift between treated and control with $p_0 = 0$ and $S = 1$, then monotonicity likely fails if the selection is predictable based on covariates
  2. This is very similar to the always-take and complier analogy
     - The bounds are exploiting the "worst case" scenarios – if the group of "shifters" are the "best" $Y(1)$, then trimming the top provides a lower bound.
     - If the "shifters" are the worst, then trimming the bottom provides the upper bound
  3. Using covariates can be used to narrow these bounds by shrinking the size of p_0

Lee Bounds (Lee 2009)

- Why is this approach great?
  - No additional instrument or model needed
  - Just monotonicity
  - Can also work in broader sample selection questions – not just wage selection, but sample attrition and other problems (recall the RD bunching approach!)

- This approach is tighter than the general Horowitz-Manski bounds (analogous to what we discussed at start of today's class)
  - Why? Monoticity restriction gets you a lot of power
  - Without that assumption, bounds are extremely wide (because the outcome is unbounded)

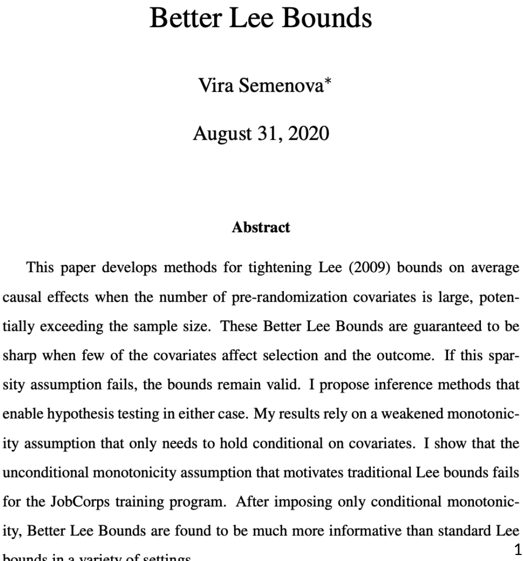
- Implemented in Stata (`leebounds`) and R (`leebounds`)

Follow-up: Better Lee Bounds (Semenova 2020)

- Makes covariate usage easier and relaxes monotonicity condition to be conditional on covariates (potentially high dimensional)
  - Lee bounds require a positive number of treated and control outcomes for each covariate value
  - Challenging with continuous values, or many

- Asymptotically sharp under “many” covariates

- Code: `https://github.com/vsemenova/leebounds`



Better Lee Bounds

Vira Semenova*

August 31, 2020

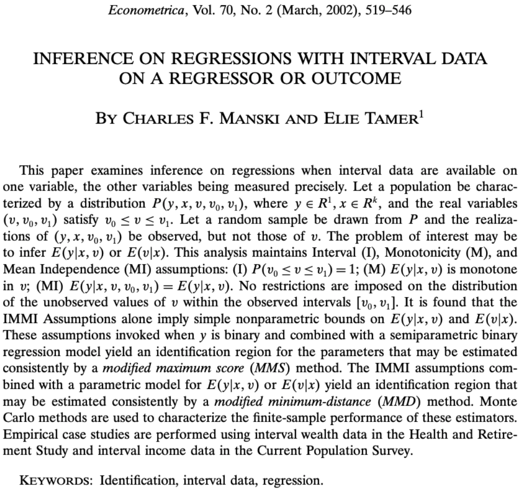
Abstract

This paper develops methods for tightening Lee (2009) bounds on average causal effects when the number of pre-randomization covariates is large, potentially exceeding the sample size. These Better Lee Bounds are guaranteed to be sharp when few of the covariates affect selection and the outcome. If this sparsity assumption fails, the bounds remain valid. I propose inference methods that enable hypothesis testing in either case. My results rely on a weakened monotonicity assumption that only needs to hold conditional on covariates. I show that the unconditional monotonicity assumption that motivates traditional Lee bounds fails for the JobCorps training program. After imposing only conditional monotonicity, Better Lee Bounds are found to be much more informative than standard Lee bounds in a variety of settings.

Manski and Tamer (2002)

- In many survey settings, data is not reported exactly, but instead in bounds
    - Wealth may be reported in ranges to encourage participation
    - Or, data may be supressed into bins to preserve anonymity (e.g. County Business Patterns data on employment)

- By definition, theis interval data should lead to set-identified parameters
    - Manski and Tamer (2002) is exactly concerned with this question



Econometrica, Vol. 70, No. 2 (March, 2002), 519–546

INFERENCE ON REGRESSIONS WITH INTERVAL DATA ON A REGRESSOR OR OUTCOME

BY CHARLES F. MANSKI AND ELIE TAMER[1]

This paper examines inference on regressions when interval data are available on one variable, the other variables being measured precisely. Let a population be characterized by a distribution $P(y, x, v, v_0, v_1)$, where $y \in R^1, x \in R^k$, and the real variables (v, v_0, v_1) satisfy $v_0 \leq v \leq v_1$. Let a random sample be drawn from P and the realizations of (y, x, v_0, v_1) be observed, but not those of v. The problem of interest may be to infer $E(y|x, v)$ or $E(v|x)$. This analysis maintains Interval (I), Monotonicity (M), and Mean Independence (MI) assumptions: (I) $P(v_0 \leq v \leq v_1) = 1$; (M) $E(y|x, v)$ is monotone in v; (MI) $E(y|x, v, v_0, v_1) = E(y|x, v)$. No restrictions are imposed on the distribution of the unobserved values of v within the observed intervals $[v_0, v_1]$. It is found that the IMMI Assumptions alone imply simple nonparametric bounds on $E(y|x, v)$ and $E(v|x)$. These assumptions invoked when y is binary and combined with a semiparametric binary regression model yield an identification region for the parameters that may be estimated consistently by a *modified maximum score* (*MMS*) method. The IMMI assumptions combined with a parametric model for $E(y|x, v)$ or $E(v|x)$ yield an identification region that may be estimated consistently by a *modified minimum-distance* (*MMD*) method. Monte Carlo methods are used to characterize the finite-sample performance of these estimators. Empirical case studies are performed using interval wealth data in the Health and Retirement Study and interval income data in the Current Population Survey.

KEYWORDS: Identification, interval data, regression.

Manski and Tamer (2002)

- Consider the data (y, v, v_0, v_1, x), where v is the true measure you'd like, which is reported with intervals $[v_0, v_1]$.

- Manski and Tamer (2002) assume:
  1. $E(Y|v, x)$ is weakly monotonic in v
  2. $E(Y|v, x, v_0, v_1) = E(Y|v, x)$

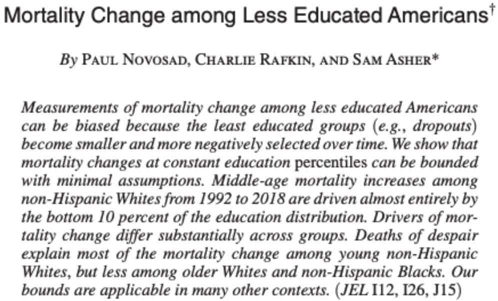
- In these settings, it is possible to put sharp bounds on the coefficients of a linear model
  - Important to note – in a multivariate regression, the set interval data on the right-hand side will also affect the coefficients for non-set interval data

- Can also be adopted to study $E(v|x)$ for covariates x
  - There are *many* data sources with type of set interval nature!

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- Mortality rates among non-Hispanic Whites without college degrees have increased substantially over time
    - Why?

- Three possible reasons:
    - an artifact of shifts in the education distribution
    - mortality could be rising uniformly among individuals in the bottom half of the education dist
    - mortality could be rising substantially at the very bottom of the education distribution



Mortality Change among Less Educated Americans[†]

By Paul Novosad, Charlie Rafkin, and Sam Asher*

Measurements of mortality change among less educated Americans can be biased because the least educated groups (e.g., dropouts) become smaller and more negatively selected over time. We show that mortality changes at constant education percentiles *can be bounded with minimal assumptions. Middle-age mortality increases among non-Hispanic Whites from 1992 to 2018 are driven almost entirely by the bottom 10 percent of the education distribution. Drivers of mortality change differ substantially across groups. Deaths of despair explain most of the mortality change among young non-Hispanic Whites, but less among older Whites and non-Hispanic Blacks. Our bounds are applicable in many other contexts.* (*JEL* I12, I26, J15)

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- Mortality rates among non-Hispanic Whites without college degrees have increased substantially over time
    - Why?

- Three possible reasons:
    - an artifact of shifts in the education distribution
    - mortality could be rising uniformly among individuals in the bottom half of the education dist
    - mortality could be rising substantially at the very bottom of the education distribution

2 *AMERICAN ECONOMIC JOURNAL: APPLIED ECONOMICS* *OCTOBER 2022*

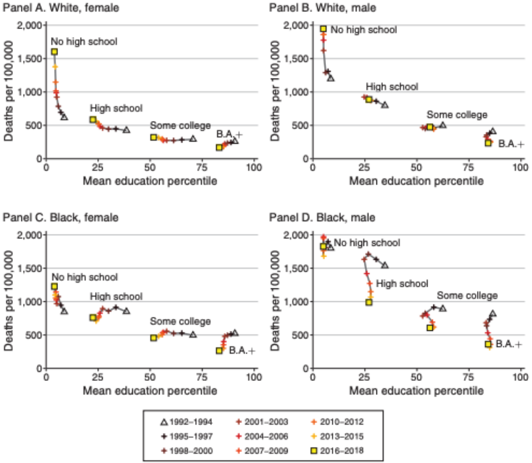


FIGURE 1. MORTALITY VERSUS EDUCATION RANK, AGE 50–54, 1992–1994 TO 2016–2018

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- This paper uses Manski and Tamer with two additional assumptions:
  1. there exists a latent education rank, which is only coarsely observed in the education data
  2. mortality rate is weakly decreasing in the latent education rank

- Effectively, put bounds on $E(y|x \in [a, b])$

2 AMERICAN ECONOMIC JOURNAL: APPLIED ECONOMICS OCTOBER 2022

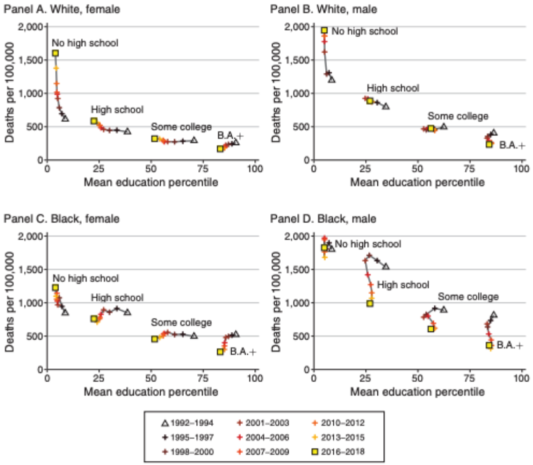


Figure 1. Mortality versus Education Rank, Age 50–54, 1992–1994 to 2016–2018

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- This paper uses Manski and Tamer with two additional assumptions:
  1. there exists a latent education rank, which is only coarsely observed in the education data
  2. mortality rate is weakly decreasing in the latent education rank

- Effectively, put bounds on $E(y|x \in [a, b])$

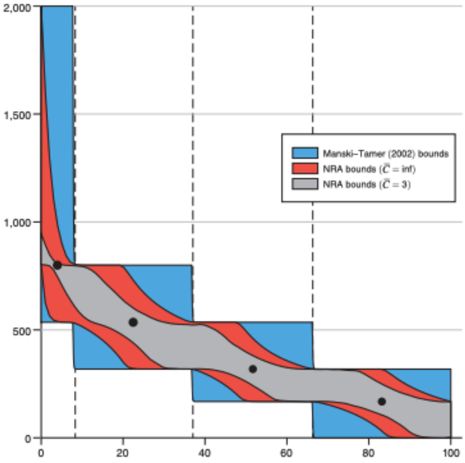


Figure 3. Change in Total Mortality of US Women, Age 50–54
Bounds on Conditional Expectation Functions

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- This paper uses Manski and Tamer with two additional assumptions:
    1. there exists a latent education rank, which is only coarsely observed in the education data
    2. mortality rate is weakly decreasing in the latent education rank

- Effectively, put bounds on $E(y|x \in [a, b])$

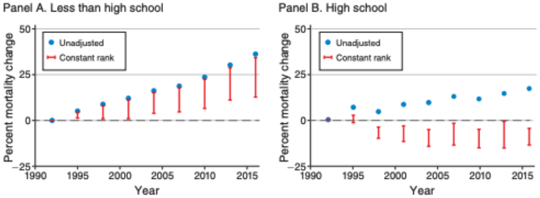


Figure 4. Changes in US Mortality, Women Age 50–54, 1992–1994 to 2016–2018: Naïve and Constant Rank Interval Estimates

Notes: Figure 4 shows mortality changes for 50–54-year-old women from 1992–1994 to 2016–2018 (all races combined), calculated under different methods. The points show unadjusted estimates for women at constant education levels—dropouts in panel A and high school graduates in panel B. Both of these population groups have shrunk as proportions of the population during the sample period. The vertical bars show bounds on mortality change in constant rank bins—ranks 0–17 in panel A and ranks 17–60 in panel B. These ranks are chosen because they are close to the share of women in 1992–1994 with less than a high school degree or exactly a high school degree, allowing the bounds to be very tight in the starting period.

Excellent application considering interval data: Novosad, Rafkin and Asher (2022)

- This paper uses Manski and Tamer with two additional assumptions:
  1. there exists a latent education rank, which is only coarsely observed in the education data
  2. mortality rate is weakly decreasing in the latent education rank

- Effectively, put bounds on $E(y|x \in [a, b])$

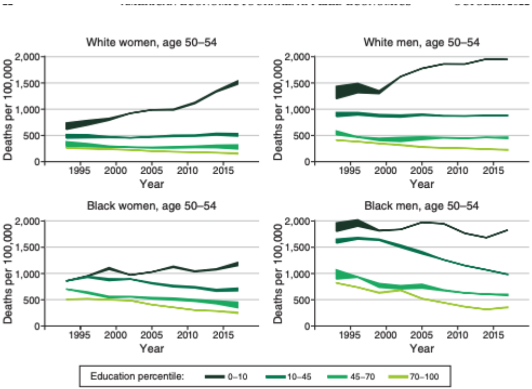


Figure 5. All-Cause Mortality Change in Constant Education Percentiles: Age 50–54, 1992–1994 to 2016–2018

Other things?

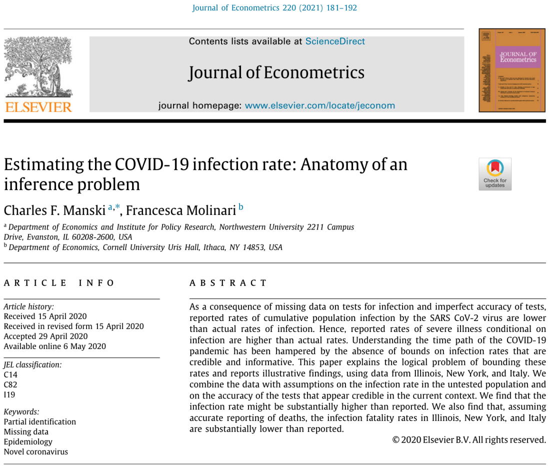
- Both of these approaches are very practical ways to deal with data issues
- They also give a hint about how to think about these applications in other settings
  - We are only scratching the surface of partial ID settings today!
- Are there other useful applications worth considering?
  - A major set related to modeling: many structural economic models imply inequalities for parameters of interest, which lead to set identification
  - What about using shape assumptions on treatments?

So why don't people use it more?

- These approaches seem quite powerful – perhaps we can still say useful things with less assumptions

- Manski and Molinari (2021) recently take this approach thinking about identifying the share of the population that has been infected with Covid-19
    - A huge issue that is plagued with a host of selection problems!

- Make some initial assumptions on the relationship between testing, symptoms and positive cases, rather than modeling the full thing in a parametric form (e.g. SIR model)



Journal of Econometrics 220 (2021) 181–192

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

ELSEVIER

Estimating the COVID-19 infection rate: Anatomy of an inference problem

Charles F. Manski [a,*], Francesca Molinari [b]

[a] Department of Economics and Institute for Policy Research, Northwestern University 2211 Campus Drive, Evanston, IL 60208-2600, USA
[b] Department of Economics, Cornell University Uris Hall, Ithaca, NY 14853, USA

ARTICLE INFO

Article history:
Received 15 April 2020
Received in revised form 15 April 2020
Accepted 29 April 2020
Available online 6 May 2020

JEL classification:
C14
C82
I19

Keywords:
Partial identification
Missing data
Epidemiology
Novel coronavirus

ABSTRACT

As a consequence of missing data on tests for infection and imperfect accuracy of tests, reported rates of cumulative population infection by the SARS CoV-2 virus are lower than actual rates of infection. Hence, reported rates of severe illness conditional on infection are higher than actual rates. Understanding the time path of the COVID-19 pandemic has been hampered by the absence of bounds on infection rates that are credible and informative. This paper explains the logical problem of bounding these rates and reports illustrative findings, using data from Illinois, New York, and Italy. We combine the data with assumptions on the infection rate in the untested population and on the accuracy of the tests that appear credible in the current context. We find that the infection rate might be substantially higher than reported. We also find that, assuming accurate reporting of deaths, the infection fatality rates in Illinois, New York, and Italy are substantially lower than reported.

So why don’t people use it more?

- Well, the problem is that the bounds are *sort of* informative, but sort of not...
  - $[0.001, 0.525]$ is a pretty wide range of infection rates
- Bit of a Rorschach test: this is either a feature or a bug
  - Reflects how strong the parametric assumptions are
  - Reflects how uninformed the policymaker is
- It is very challenging to present bounds that are this uninformative in resarch papers
  - Good goal is as a supplement
  - Less true in IO style settings!

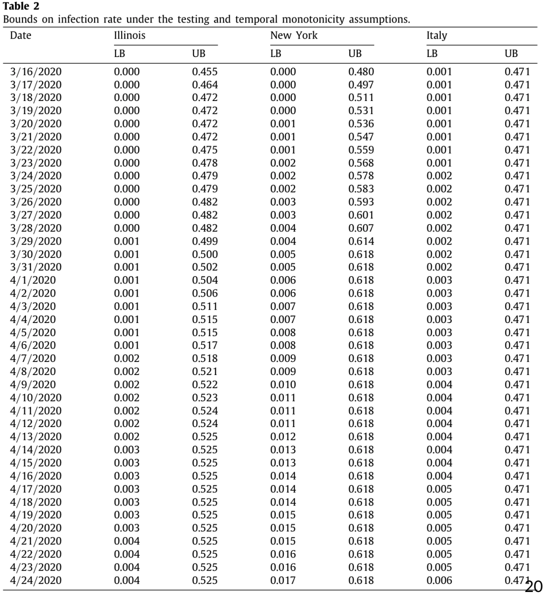


Table 2
Bounds on infection rate under the testing and temporal monotonicity assumptions.

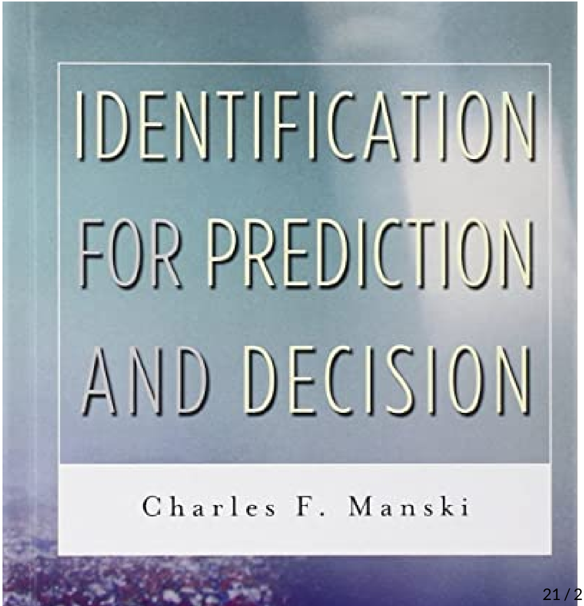
Date	Illinois		New York		Italy	
	LB	UB	LB	UB	LB	UB
3/16/2020	0.000	0.455	0.000	0.480	0.001	0.471
3/17/2020	0.000	0.464	0.000	0.497	0.001	0.471
3/18/2020	0.000	0.472	0.000	0.511	0.001	0.471
3/19/2020	0.000	0.472	0.000	0.531	0.001	0.471
3/20/2020	0.000	0.472	0.001	0.536	0.001	0.471
3/21/2020	0.000	0.472	0.001	0.547	0.001	0.471
3/22/2020	0.000	0.475	0.001	0.559	0.001	0.471
3/23/2020	0.000	0.478	0.002	0.568	0.001	0.471
3/24/2020	0.000	0.479	0.002	0.578	0.002	0.471
3/25/2020	0.000	0.479	0.002	0.583	0.002	0.471
3/26/2020	0.000	0.482	0.003	0.593	0.002	0.471
3/27/2020	0.000	0.482	0.003	0.601	0.002	0.471
3/28/2020	0.000	0.482	0.004	0.607	0.002	0.471
3/29/2020	0.001	0.499	0.004	0.614	0.002	0.471
3/30/2020	0.001	0.500	0.005	0.618	0.002	0.471
3/31/2020	0.001	0.502	0.005	0.618	0.002	0.471
4/1/2020	0.001	0.504	0.006	0.618	0.003	0.471
4/2/2020	0.001	0.506	0.006	0.618	0.003	0.471
4/3/2020	0.001	0.511	0.007	0.618	0.003	0.471
4/4/2020	0.001	0.515	0.007	0.618	0.003	0.471
4/5/2020	0.001	0.515	0.008	0.618	0.003	0.471
4/6/2020	0.001	0.517	0.008	0.618	0.003	0.471
4/7/2020	0.002	0.518	0.009	0.618	0.003	0.471
4/8/2020	0.002	0.521	0.009	0.618	0.003	0.471
4/9/2020	0.002	0.522	0.010	0.618	0.004	0.471
4/10/2020	0.002	0.523	0.011	0.618	0.004	0.471
4/11/2020	0.002	0.524	0.011	0.618	0.004	0.471
4/12/2020	0.002	0.524	0.011	0.618	0.004	0.471
4/13/2020	0.002	0.525	0.012	0.618	0.004	0.471
4/14/2020	0.003	0.525	0.013	0.618	0.004	0.471
4/15/2020	0.003	0.525	0.013	0.618	0.004	0.471
4/16/2020	0.003	0.525	0.014	0.618	0.004	0.471
4/17/2020	0.003	0.525	0.014	0.618	0.005	0.471
4/18/2020	0.003	0.525	0.014	0.618	0.005	0.471
4/19/2020	0.003	0.525	0.015	0.618	0.005	0.471
4/20/2020	0.003	0.525	0.015	0.618	0.005	0.471
4/21/2020	0.004	0.525	0.015	0.618	0.005	0.471
4/22/2020	0.004	0.525	0.016	0.618	0.005	0.471
4/23/2020	0.004	0.525	0.016	0.618	0.005	0.471
4/24/2020	0.004	0.525	0.017	0.618	0.006	0.471

Recommended further reading

- “Identification for Prediction and Decision” (Manski 2007)

- “Partial Identification of Local Average Treatment Effects With an Invalid Instrument” Flores and Flores-Lagunes (2013)

- “Estimation and Confidence Regions for Parameter Sets in Econometric Models” Chernozhukov, Hong and Tamer (2007)



IDENTIFICATION FOR PREDICTION AND DECISION

Charles F. Manski