ID: 12174137

Program Evaluation Final

12/7/2018

1. The fundamental problem in trying to estimate the effect of incarceration on employment or recidivism is that we only observe one of the two potential outcomes for each individual in the studies. For example, in Harding et al, we only observe either a positive employment outcome for an individual or a negative employment outcome for an individual, while in Hjalmarsson we either see recidivism or not for an individual. Therefore, we have to estimate the missing counterfactual in order to determine the impact for any given individual.
2. a) Instrument relevance assumes that the instrument you have chosen is correlated with the causal variable that you’re interested in. Instrument validity assumes that you have instrument relevance as well as exogeneity, which means that the instrument you have chosen is only correlated with the outcome variable through the treatment variable and not through any unobservables. Lastly, monotonicity assumes that the instrumental variable will move all individuals in the sample in the same direction – there are no ‘defiers’ in the sample.

b) Within the scope of the Harding et al setting, judge identifiers serve as instruments for the sentence type (either probation or prison). In this scenario, instrument relevance requires that the judge identifier be correlated with the sentence type that’s handed down to the individual. Instrument validity requires instrument relevance and exogeneity, which in this scenario means that the labor outcomes in this study are only impacted by the sentence received by the individual and not by the judge that sentenced said individual, which implies a random assignment of judges to cases. Monotonicity in this study requires that ‘harsh’ judges always impose harsher sentences than their colleagues, as opposed to imposing harsher sentences in some cases and more lenient sentences in other cases.

1. Harding et al claim that the instrument relevance assumption is met by analyzing how the probability of sentencing to probation or prison varies by judge within county. In this analysis, they find clearly evident judge variation from the mean in the probability of a prison vs a probation sentence, in prison minimum sentence lengths, and in probation sentence lengths. This suggests that judges have different levels of leniency which are clearly correlated to the treatment; a judge whose harshness is far away from the mean will give a sentence of prison all else equal, while a judge whose leniency is far away from the mean will give a sentence of probation all else equal. In other words, the instrumental variable is clearly correlated to the treatment: probation or prison.

Harding et al also test the strength of the correlation, since variations in the treatment across instruments could signal only a weak correlation between instrument and treatment. They test it by pulling out Shea’s partial R-squared metric to show the proportion of the variation in each treatment variable independently explained by instruments. This then shows that the first-stage equation explains a substantial portion of the variation across all treatments and subgroups, which they claim gives proof of a strong instrument. They also use the F-test, which gives them mostly values greater than ten for large sample sizes, yet smaller values for smaller samples. Lastly, they test for over-fitting by re-estimating their models on fewer instruments and found estimates of the same direction and magnitude as their preferred model, thus suggesting no over-fitting.

On the whole, I’m convinced that the assumption is satisfied. The judges clearly have different levels of leniency which are correlated to the treatment variable: harsher judges give more prison time and more lenient judges give more probation. While the F-test fails for smaller sample sizes, it still is above the ideal number of 10 for the overall sample. This is complemented by partial R-squared metrics that show that the first-stage equation explains a substantial portion of the variation across treatments, and a test that shows that there is no over-fitting in this analysis.

1. Harding et al claim the instrument validity assumption is met through an understanding of how Michigan judges are assigned to cases. As they explain how Michigan judges are assigned to cases, it becomes evident fairly quickly that the assignment is completely random and that there’s little opportunity to circumvent the process and little motivation to do so for minor cases. They also check to make sure that the covariates they observe are uncorrelated with judge assignment by conducting a joint significance test which finds small F-values that are statistically significant and thus indicate that there is no correlation between covariates and judge assignment.

For the assumption to hold, Harding et al also check to see if the instrumental variable is only correlated to the outcome via the treatment. In proving that, they rely on reasoning through an example where a defendant would return to a life of crime if treated harshly in the courtroom by a harsh judge and thus lose confidence in the judicial system. Lastly, they check for a potential violation of this correlation in the assignment of pretrial detention and find only a slight correlation conditional on race and work history. They also check for prosecutor reaction to being assigned a harsh or lenient judge, and for defendant plea bargaining depending on the judge assigned, where they find no effect by analyzing if lenient judges have more trials and fewer plea bargains than the rest of the judges in the data.

On the whole, I’m convinced this assumption is satisfied. It’s made evidently clear that Michigan judges are randomly assigned to cases, which means that labor outcomes are impacted by the sentence levied by the judge and not by the selection of a judge for the case. While the check on the exclusivity condition – that the only correlation between a labor outcome and the judge is through the sentence – isn’t explored fully, most of the evidence (the contrived example of a defendant turning to a life crime due to being treated harshly in court, the small correlation in assignment to pretrial detention by a judge, and no evidence on defendants seeking plea bargains or prosecutors reacting to the judges assigned) seems to indicate that it’s satisfied.

1. Instead of attempting to meet the monotonicity assumption, Harding et al relax it by interacting judge dummies with presentencing individual characteristics and also treating those interactions as instruments. Doing so, the authors claim, allows them to identify LATE on compliers, defiers, and both populations jointly.

However, I’m not entirely satisfied by that explanation. While there’s evidence that you can relax the monotonicity assumption in this way in order to interpret local average treatment effects on different groups, it’s still unclear if the new instrumental variables you created by interacting judge dummies with presentencing individual characteristics meet the assumptions of instrument relevance and instrument validation. In order to be satisfied that the monotonicity assumption is met -- or more precisely, in order to be satisfied that relaxing monotonicity in this scenario is appropriate, I’d like to see evidence that the new instrumental variables created satisfy all the relevant IV assumptions outlined above.

1. The estimates in terms of time since sentencing are likely to be more valid than the estimates in terms of times since release for a number of reasons. To begin, estimating in terms of time since sentencing allows you to measure the total effect of imprisonment, which combines the incapacitation effect of serving time in prison versus being on probation and the effects that persist after the release of the prisoner. What’s more, starting the estimate at sentencing allows you to avoid issues of time. If you were to measure the estimate post-release, you could potentially have differences in age between prisoners and people on probation. Since you would begin to track people sentenced to probation immediately after sentencing, and people sentenced to prison after their release, it’s highly likely that your population of people sentenced to prison would be older at the start of the risk period due to the time they spent in prison. This could bias your estimate if, for example, older people have a higher likelihood of finding work than younger people.
2. The before-after estimator assumes that there is no counterfactual trend, which means that labor market outcomes for people sentenced to prison would have remained constant in the absence of a sentence to prison, that is they would have been the same as the outcomes for those sentenced to probation. In order to make this assumption valid, it’s also important to assume that there is no selection bias between the prison and probation groups. The data does not back this assertion, given that labor outcomes for people sentenced to prison are different from those of people sentenced to probation prior to sentencing, and certainly afterwards as well.
3. The authors could have implemented a differences-in-differences approach if they treated people sentenced to probation as a control group and people sentenced to jail as the treatment group, with the sentence being the naturally occurring event that creates a shift for the treatment group and not the control group. In order to implement a DiD, the authors would have to assume parallel trends between the treatment and control group, which in this case means that labor trends for the treatment and control groups would have been not necessarily equal, but parallel in their trendlines prior to sentencing. In looking at Fig. 1 and Fig. 2, it’s clear that the parallel trends assumption is not met due to the fact that there’s a sharp drop-off in employment prior to sentencing for people who are about to be sentenced to prison, and no similar drop-off immediately prior to sentencing for those about to be sentenced to probation. That said -- assuming I want to account for pre-trail detention to get the full incapacitation effect -- if I were to estimate the effects of incarceration on employment 24 quarters after sentencing for blacks and whites, I’d estimate that the effect of incarceration on employment 24 quarters after sentencing for blacks would be the percentage of blacks employed 24 quarters after a probation sentence minus the percentage of blacks employed at probation sentencing, all minus the percentage of blacks employed 24 quarters after a prison sentence minus the percentage of blacks employed at prison sentencing. This would be (.19 - .31) - (.09 - .05) = **-0.16**, or a reduction of 16% in employment outcomes for blacks if incarcerated as opposed to put on probation. For whites, this same reasoning would lead to (.31 - .46) – (.15 - .08) = **-0.22**, or a reduction of 22% in employment outcomes for whites if incarcerated as opposed to put on probation.
4. For a regression discontinuity to identify the effect of incarceration, several assumptions have to be satisfied. The crucial assumption is the identifying assumption which in this case means that you’d want the unobservable characteristics of youth to not vary immediately around the sentencing cutoffs set by the 1986 guidelines. Another assumption is that your treatment varies discontinuously around the cutoff, in this case that there’s discontinuous variation around incarceration around the cutoffs prescribed by the sentencing guidelines. Lastly, you want the running variable to not be manipulable, which in this case means that neither juveniles nor courts can choose whether a juvenile is incarcerated or not.
5. In support of the identifying assumption, Hjalmarsson graphs different aspects of the data around the sentencing cutoffs. She holds current offense class constant and plots the percentage of individuals incarcerated for each adjudication score in that row. She finds no discontinuities in treatment for the offense classes that are the most and least grave (B and D, respectively), and finds discontinuities in treatment for all the other offense classes. She also holds the rounded adjudication score constant and plots the percentage of individuals incarcerated in each cell of that columns, and finds clear discontinuities in treatment.

In order to test whether there are no discontinuities in unobservables around the cutoff, which is something that can’t be observed the directly, the author chooses to examine if observable variables that correlate with recidivism are continuous around the cutoff. In order to test for this condition, she estimates a Cox proportional hazard model to estimate the hazard ratio. She then regresses that on the treatment variable, the adjudication score, and the current offense class dummies and finds that there’s balance in these observable characteristics around the cutoff.

Finally, in testing for manipulation, she checks for gaming on part of the courts and of the juveniles. The identifying assumption would not be satisfied if prosecutors or judges could use discretion charge juveniles with crimes that place them on either side of the cutoff, or juveniles could control their adjudication score. To assess whether prosecutors could be gaming the guidelines, the author determines that a prosecutor might be incentivized to game the system by making prosecutorial decisions today based on the juvenile’s future score, which is determined to be highly unlikely since the 1998 sentencing reform changed the guidelines so fundamentally that prosecutorial discretion could not have played a role in whether criminal history scores fall on either side of the cutoff. Meanwhile judges would be incentivized to sentence based on the treatment that might have a greater effect on the juvenile but have no hand in calculating scores which minimizes their ability to game the system and use their discretion in sentencing outside of the guidelines very infrequently. Lastly, in order for juveniles to game the system, they would have to have prior knowledge of the new sentencing grid, a proposition that the author finds a preposterous given how poorly publicized the change in sentencing guidelines was. In the case of prosecutors and juveniles, the author also tests for balance of observables around the cutoff for prosecutors to identify and discontinuities that might indicate manipulation, as well as the probability of juveniles committing an offense that would place them in the class immediately below the cutoff. In both cases, she finds no empirical evidence on manipulation.

The evidence presented in support of these assumptions is convincing. The graphs that test for discontinuity in the treatment around the cutoff and continuity in observables around the cutoff show clear evidence that both of those conditions hold for the caliper being analyzed. Secondly, although it’s impossible to observe if unobservable characteristics vary around the cutoff, the proxy measure of observing variation of observable characteristics around the cutoff finds evidence of clear balance in these characteristics which we can extrapolate to balance among the unobservables. Lastly, it’s clear from her explanations that any intent to manipulate might be prohibitively hard to undertake, and her empirical analysis bears that out given the significance of the tests as determine by the R-squared values in the test of balance for prosecutors and t-test statistics in the empirical analysis of juvenile manipulation.

1. The author’s analysis initially lends itself to a fuzzy RD since within the sentencing guidelines, there are a few exceptions to the otherwise strict rules set out for sentencing: judges can depart from the standard range of guidelines in the case of Chemical Dependency Disposition Alternative cases, Special Sex Offender Disposition Alternative cases, and if the judge believes the standard sentence is too harsh or lenient, they can decrease or increase the severity of the sentence. However, she chooses to drop the cases of the 738 youth who do not follow the assignment mechanism; in other words, the youth whose judges gave them a more lenient or harsh sentence that what the guidelines would indicate. By doing so, she turns her analysis into a sharp RD since now treatment is deterministic. Had she kept the fuzzy RD, her treatment parameter would have intent-to-treat, but by dropping the defiers, the treatment parameter becomes a LATE since she only obtains an internally valid estimate for a very strictly defined segment of the population. The lack of external validity in her study design prevents her from measuring average treatment effects beyond the caliper.
2. I would turn the regression model into a linear model. The current regression estimate seems to indicate continuity in outcomes for the probation group in Fig 1 and Fig 2 when in reality there is none. A linear model, especially within a smaller caliper window, would more clearly capture a discontinuity at the point of treatment.

Mario Moreno