Identification in the two Chetty papers

Nadav Kunievsky

March 3, 2021

In a series of papers, Chetty and co-authors, takes a different approach to estimate the role neighborhoods play in shaping income outcomes. Chetty and his co-autors lay out their argument in three acts. Chetty et al. (2014), show some suggestive evidence implying that different Commuting Zones (CZ) and counties are associated with different levels of social mobility, as measured by their rank-rank regression. In Chetty and Hendren (2018a) they examine the causal effect of growing up in a better neighborhood, as measured by the average income rank of the permanent residents. In this paper, they also establish that the effect is almost linear in exposure time. Chetty and Hendren (2018b), and Chetty and Porter (2018) build on the intuition in Chetty and Hendren (2018a) and estimate a linear model for each CZ/county/Census Tract in the US, to create a mobility map, for different locations.

1 Identification in Chetty and Hendren (2018a)

Chetty and Hendren (2018a), use federal income tax record spanning 1996–2012, on individuals born between 1980-1988, to measure the causal effect of living in a better neighborhood, as measured by the the average income of the permanent residents' children, on the long terms outcomes of the children of families who moved to those neighborhoods. Using a semi-parametric specification, they find an almost linear effect of living in a better neighborhood on the child percentile in the income distribution. Specifically, they find that, *"on average, spending an additional year in a CZ where the mean income rank of children of permanent residents (PR) is 1 percentile higher (at a given level of parental income) increases a child's income rank in adulthood by approximately 0.04 percentiles. That is, the incomes of children who move converge to the incomes of PR in the destination at a rate of 4% per year of childhood exposure"*.

To get this result, Chetty and Hendren (2018a) start by defining neighborhood quality, for parents at percentile p of the income distribution, by examining the expected income percentile of children of PR , given their parents income percentile. To do so they run the following regression on children of non-movers

$$y_i = \alpha_{cs} + \psi_{cs} p_i + \varepsilon_i$$

where y_i is the percentile of the income of children at the age of 24, α_{cs} is fixed effect for CZ, c, and cohort s, p_i is the the parents percentile in the income distribution and ε_i is the error term. Then, they take the predicted value, $\bar{y}_{pcs}$, and use it as proxy for the neighborhood quality.

The authors then move on to examine the causal effect of increasing this quality measure on children outcomes. Consider the following outcome model for a child who moves at the age of m

$$y_i = \alpha_{o,m} + \beta_m \bar{y}_{pds} + \theta_i$$

where y_i is child i outcome, m is the age the parent moved to a new neighborhood, d, $\bar{y}_{pds}$ is the the quality of the new neighborhood d, for parent of percentile p and child from cohort s. $\alpha_{o,m)}$ is the fixed effect of of moving out of neighborhood o at the age of m, and θ_i is the error term. Typically, the error term and the neighborhood quality are going to be correlated. Specifically, let b_m be the coefficient from the regression model, then the bias of the estimator is given by

$$b_m = \beta_m + \frac{\mathrm{Cov}(\theta_i, \bar{y}_{pds}|m)}{\mathrm{Var}(\bar{y}_{pds}|m)}$$

Although Chetty and Hendren, 2018a can't identify β_m directly from the data, they can estimate what they call the exposure effect, $\gamma_m = \beta_m - \beta_{m+1}$, by imposing the assumption[1] that the selection effect does not vary with the child's age at the move, i.e. $\frac{\mathrm{Cov}(\theta_i, \bar{y}_{pds}|m)}{\mathrm{Var}(\bar{y}_{pds}|m)}, = \frac{\mathrm{Cov}(\theta_i, \bar{y}_{pds})}{\mathrm{Var}(\bar{y}_{pds})}, \forall m$. With this assumption that exposure effect is just the difference between the regression coefficient of two regressions for movers at the age of m and for movers at the age of $m+1$, $\gamma_m = \beta_m - \beta_{m+1} = b_m - b_{m+1}$.[2] In practice, Chetty and Hendren, 2018a estimate

[1]Their model also impose two additional assumptions, first, that the neighborhood effect is linear and symmetric in the PR children outcomes, and different neighborhoods with the same quality level, are equivalent

[2]We can also think of their identification as a DID setup. Assume that given the origin neighborhood and the destination, there is no selection on the age of the moving, i.e. $E(\theta_i|m, o, \bar{y}_{pds}) = E(\theta_i|o, \bar{y}_{pds})$, then we can identify the exposure effect by taking the difference between the differences

$$E(y_i|m, o, \bar{y}_{pds}) - E(y_i|m+1, o, \bar{y}_{pds}) = \alpha_{o,m} - \alpha_{o,m+1} + (\beta_m - \beta_{m+1})\bar{y}_{pds}$$
$$E(y_i|m, o, \bar{y}_{pds} + 1) - E(y_i|m+1, o, \bar{y}_{pds} + 1) = \alpha_{o,m} - \alpha_{o,m+1} + (\beta_m - \beta_{m+1})(\bar{y}_{pds} + 1)$$

where we used the timing assumption to remove the error terms.

the following semi-parametric regression model[3]

$$y_i = \alpha_{qosm} + \sum_{m=9}^{30} b_m I\left(m_i = m\right)\Delta_{odps} + \sum_{s=1980}^{1987} \kappa_s I\left(s_i = s\right)\Delta_{odps} + \varepsilon_{2i}$$

where y_i is the movers child outcome, α_{qosm} is CZ by parent income decile by birth cohort by age at move fixed effect, I is an indicator function, and $\Delta_{odps} = \bar{y}_{pds} - \bar{y}_{pos}$[4]

In addition to the main specification, Chetty and Hendren, 2018a, they validate the assumption using a series of tests, which focus on selection. All the tests implies that their assumption on selection holds

1. Comparing siblings, by running the main specification with family fixed effect, and control for family time-varying observables (parents income and martial status).

2. They identify instances of outflow shocks, where they observe a large out migration from a ZIP code. They assume that these displacement shocks are exogenous, and use them as instruments in the main specification.

3. They preform placebo tests, in which they replace $\bar{y}_{pds}$ with the average quality based on the children of PR from different cohort, s', from $\bar{y}_{pds'}$ and find almost zero effect.

2 Chetty and Hendren (2018b)

Motivated by the results of the previous paper and implied linear effect of the neighborhood effect, Chetty and Hendren (2018b), aim to estimate the causal effect of each CZ and county in the US, using administrative IRS data, where the main specification uses the same sample as Chetty and Hendren (2018a). To estimate the causal effect of each CZ/county they start by assuming the following statistical model[5]

$$y_i = \alpha_{od} + \boldsymbol{e}_i \cdot \boldsymbol{\mu} + \epsilon_i$$

[3]They also run a more restrictive version of this regression model, which imposes linearity in the effect and in the parent's income percentile

[4]Notice that as the authors control for the location fixed effect, all the variation in Δ_{odps} comes from $\bar{y}_{pds}$. The author define this explaining variable as such for ease in interpretating of the results.

[5]In practice, they estimate a more flexible model, which allows for different effects of children with different parental ranking.

$$y_i = \alpha_{od} + \alpha_{od}^{P} p + \boldsymbol{e}_i \cdot \boldsymbol{\mu}_p + g_{od}\left(p_i, s_i\right) + \varepsilon_i$$

where α_{od}^{P} is interaction of the fixed effect with the parental income ranking, $g_{od}\left(p_i, s_i\right)$ is a flexible function to control for the cohort effect and parental ranking, and $\boldsymbol{\mu}_p$ is a vector with $\{m_{pc}\}$, where $m_{pc} = \mu_c + \mu_1 p$.

where α_{od} is a fixed effect of origin-destination. $\boldsymbol{e}_i = \{e_{ic}\}$ is a vector whose entries denote the number of years of exposure that child i has to place c, before age A, and $\boldsymbol{\mu}$ is a vector of coefficients that capture CZ exposure effect. To identify the exposure parameters, Chetty and Hendren (2018b) again impose their timing assumption, where, conditional of α_{od} the exposure time to each place $\boldsymbol{e}_i$ is orthogonal to other determinants of children's outcomes, i.e, $\text{Cov}(e_{ic}, \epsilon_i) = 0$ for all c[6]. Estimating the model produces estimates for the causal effect of staying a one additional year in a neighborhood.[7]. Unfortunately, some of these estimates are quite noisy as there aren't a lot of movers. Therefore, they use a Bayesian shrinkage procedure to reduce MSE.

2.1 Results

Chetty and Hendren (2018b) focuses on the causal effect of neighborhoods on families at the 25th percentile and 75th percentile of income distribution. They find that there's some dispersion in effects, across CZ and counties (The causal effect of an increase in one standard deviation in the causal effect parameter, for families at the 25 percentile, increases the child future ranking by 0.17, per year). They find that the neighborhood effect is more important for boys than girls. They then move and use the estimates to explore correlation between different neighborhood characteristics and the neighborhood effect, and find that the causal that segregation and inequality are negatively correlated with the effect, where school quality is positively correlated.

References

Chetty, Raj, J. N. F. N. H. M. R. J. and Porter, S. R. (2018). The opportunity atlas: Mapping the childhood roots of social mobility. *National Bureau of Economic Research,No. w25147*.

Chetty, R. and Hendren, N. (2018a). The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162.

[6]Following footnote 3, we can think of the identification of taking differences between the following two moments

$$E(y_i|o, d, m) = \alpha_{od} + m\mu_o + (A - m)\mu_d + E(\epsilon_i|o, d, m)$$
$$E(y_i|o, d, m + 1) = \alpha_{od} + (m + 1)\mu_o + (A - (m + 1))\mu_d + E(\epsilon_i|o, d, m)$$

which would give the $\mu_d - \mu_o$. Notice that we are implicitly assuming here a constant effect, that is the same whether we migrate from o to d, or from d to o, fixed by age of moving, and also the same whether someone moving between o and d or whether move between o and d'.

[7]Chetty and Hendren (2018b) normalize the effect to be relative to the mean effect

Chetty, R. and Hendren, N. (2018b). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Chetty, R., Hendren, N., Kline, P., Saez, E., , and Turner, N. (2014). Is the united states still a land of opportunity? recent trends in intergenerational mobility. *American Economic Review*, 104(5):47–141.