Intergenerational Mobility By Race with Limited Data

Jeremy Majerovitz

February 2, 2018

Introduction

What determines children's eventual outcomes as adults? An extensive literature has examined the role of parent's income in determining the outcomes of their children (Black and Devereux, 2011; Chetty et al., 2014). However, there is relatively less work looking at the effect of race. Although some have examined this question (Bhattacharya and Mazumder, 2011; Hertz, 2005; Isaacs, 2007; Kearney, 2006; Mazumder, 2014), this work is made difficult by limited data. Estimating intergenerational mobility typically involves a data set that links parents to children and contains contains data on income in adulthood for both; estimating intergenerational mobility by race typically involves using a data set like this which also contains information on race. Data quality, sample size, and attrition are serious concerns in estimating intergenerational mobility, and the literature has benefitted greatly from recent work using large, high-quality administrative data, notably (Chetty et al., 2014) who use data from the IRS linking parents and children. However, the tax data do not contain information on race, and the longitudinal surveys that the literature has thus relied on (e.g. PSID, NLSY, SIPP) suffer heavily from the aforementioned problems, and sample size concerns are exacerbated when attempting to obtain precise estimates for racial subgroups.

This paper focuses on how to estimate intergenerational mobility by race given these data limitations. I combine moments from multiple high quality data sets, each of which provides a pairwise linkage between two of the three variables: race, parent income, and child income.[1] The estimator I use is derived from an inversion of the omitted variables bias formula, and requires no additional statistical assumptions beyond the standard OLS setup.

It is difficult to overstate the importance of intergenerational mobility by race in the United States. A large and persistent black-white gap in outcomes has been one of the, and perhaps *the*, defining issues in American economic history. Leaving aside the important issue of whether income is a good proxy for the

[1] In the future, I plan to also explore other methods as well; I discuss these in the conclusion.

wide range of social and economic outcomes that differ across racial groups (or more generally a proxy for welfare), estimating intergenerational mobility by race can answer three important questions related to the black-white gap. First, how much does race matter for children's success, once we control for the effect of parent's income? This is not obvious from the black-white gap alone: since black parents are poorer than white parents, some of the black-white gap is the residual effect of past disadvantage rather than the contemporary effects of race. Second, previous work has suggested that parent income plays a sizable role in determining child outcomes: how much does parent's income still matter once we account for race? Earlier work may suffer from a form of omitted variables bias; e.g. poorer parents are more likely to be black, and assuming black children face income-lowering discrimination as adults, this will assign to parent income an effect that is really due to race.

Third, where is the black-white gap headed? Understanding the relationship between parent income, child income, and race allows us to describe the steady-state of the black-white gap. How much should we expect the gap to narrow, or are we already at the steady-state? Of course, this assumes that intergenerational mobility by race is stable over time: the next stage of this project is to compute estimates for many cohorts and see how integenerational mobility by race evolves over time.

The paper proceeds as follows. The next section describes the econometric problem I face and derives the estimator I will use to solve this problem. The following section describes the data in more detail and presents results. In the final section, I conclude by discussing the results and next steps.

Econometric Problem and Solution

To estimate intergenerational mobility by race, we must first formalize this concept. Since past work has typically focused on regressions of child income on parent income, a natural approach is to consider a model regressing child income on parent income and race. Formally, we will focus on the model:

$$Y = X\beta + W\delta + \varepsilon$$

where Y is child income percentile, X is parent income percentile (and a vector of ones), and W is a set of indicators for race.[2] We impose standard OLS assumptions, namely we identify the model with $E[X'\varepsilon] = 0$ and $E[W'\varepsilon] = 0$.

The natural way to estimate this and related models is using a data set that contains all three sets of variables (race, parent income, and child income), and simply running a regression. Some papers have taken

[2]Here, I include Hispanic status as one of the racial indicators. Although preferred terminology differs on whether to refer to Hispanics as a race or some other type of category, I believe Hispanic status is important to include in my analysis regardless.

this approach (Bhattacharya and Mazumder, 2011; Hertz, 2005; Isaacs, 2007; Kearney, 2006; Mazumder, 2014). However, the precision and reliability of these results is hampered by the fact that the longitudinal surveys they use are not especially large or high-quality data sets (by quality, I refer both to concerns about the accuracy of the income data and especially to concerns about selective attrition and a non-representative sample). The use of tax data by Chetty et al. (2014) introduced a large, high-quality data set to the literature, but the tax data does not have information on race.

Although we cannot access a large, high-quality data set describing the joint distribution of race, parent income, and child income, we can easily access large, high-quality data on the pairwise distributions of each of these variables. The tax data provides the link between parent and child income, and the necessary statistics have been made publicly available by Chetty et al. (2014) and Chetty and Hendren (2017). Parent and child income distributions by race can be obtained from different years of the Census. I will thus show how to estimate our model using only moments gathered from pairwise distributions.

First, note the omitted variables bias formula:

$$\beta = E\left[\hat{\beta} - \eta_{WX}\delta\right]$$
$$\delta = E\left[\hat{\delta} - \eta_{XW}\beta\right]$$

where $\hat{\beta}$ and $\hat{\delta}$ denote the naive OLS estimators (e.g. $\hat{\beta} = [X'X]^{-1}X'Y$) and η_{WX} and η_{XW} are the regression coefficients of W on X and of X on W respectively (e.g. $\eta_{WX} = [X'X]^{-1}X'W$). Solving the system of equations and replacing expectations with probability limits[3] we get:

$$\beta = \text{plim}\left(I - \eta_{WX}\eta_{XW}\right)^{-1}\left(\hat{\beta} - \eta_{WX}\hat{\delta}\right)$$
$$\delta = \text{plim}\left(I - \eta_{XW}\eta_{WX}\right)^{-1}\left(\hat{\delta} - \eta_{XW}\hat{\beta}\right)$$

This formula depends solely on pairwise moments, rather than on the full joint distribution, and thus we can estimate this in our setting. Two points are worth emphasizing. First, this formula is actually just another way of writing the standard OLS formula; it just makes clear that the OLS formula is a function of a set of pairwise moments. If we applied this estimator to a single data set which had the joint distribution of Y, X, and W, the results would be numerically identical to OLS. Second, this strategy will not allow for the estimation of interaction effects between X and W (in our setting this is different slopes by race), since

[3] If we had everything in one data set then we could use the expectation operator, but since we have different data sets we will rely on consistency instead.

we do not observe $X \otimes W$ in the same data set as we observe Y (this would require a data set with the joint distribution).

Standard Errors

One issue I am still working on is calculating and computing the standard errors of this estimator. To begin, it is useful to note that the problem can be set up as a GMM problem with the following score function:

$$g(Z_i, \theta_i) = \begin{pmatrix} \beta - \hat{\beta} + \eta_{WX}\delta \\ \delta - \hat{\delta} + \eta_{XW}\beta \\ (W_i' - X_i'\eta_{WX})X_i \mathbf{1}_{\text{Parent Census}} \\ (X_i' - W_i'\eta_{XW})W_i \mathbf{1}_{\text{Parent Census}} \\ (Y_i' - X_i'\hat{\beta})X_i \mathbf{1}_{\text{IRS}} \\ (Y_i' - W_i'\hat{\delta})W_i \mathbf{1}_{\text{Child Census}} \end{pmatrix}$$

where $\mathbf{1}_{\text{Parent Census}}$ is an indicator that the observation is contained in the parent sample in the Census, $\mathbf{1}_{\text{Child Census}}$ indicates being in the child sample, and $\mathbf{1}_{\text{IRS}}$ indicates being in the IRS sample. Being able to set the problem up as a GMM problem is useful because it tells us that the estimator is consistent and asymptotically normal, and in principle gives us a formula for how to compute the standard errors. Although we cannot tell when an observation from one sample is also in another, we do know that the sampling is random and we know the probability that, for example, an observation observed in the parent sample will be observed again in the child sample (ignoring the possibility of early mortality). I am still, however, working out the details of this issue, and thus I have not yet computed standard errors for my estimates.

Heterogeneous Slopes

Another issue I am still working on is how to estimate a model with heterogeneous slopes by race. Although this cannot be done with the simple OLS assumption that the error term is orthogonal to the regressors, this can be achieved if we strengthen the assumption from orthogonality to conditional mean independence (i.e. $E[\varepsilon|X, W] = 0$). Intuitively, if we assume linearity, then any non-linearities in the relationship (not conditional on race) between parent percentile and mean child percentile must come from differing slopes by race, and the fact that the distribution of races is not uniform over income. This also suggests the potential difficulty of this approach: pinning identification on linearity is essentially identification off of functional form, and is thus not appealing.

More formally, we can gain identification by adding moment conditions of the form:

$$E[(Y_i - W_i'\delta - \beta X_i - \gamma W_i' \otimes X_i')f(X_i)] = 0$$

where $f(X_i)$ is a technical instrument, that is a function solely of X_i. Expanding this moment condition out, it is clear that this moment condition is composed of a number of moments which can each be estimated in different parts of the data (nothing relies on observing Y and $X \otimes W$ in the same data). However, I have not worked out the details and implemented this, and I am unsure of whether it is worth it given how heavily it relies on the unsatisfying assumption of linearity.

Results

To implement this strategy, I need data on the pairwise distributions of parent income, child income, and race. For the linkage between parent and child income, I use statistics generated from the tax data by Chetty et al. (2014) and Chetty and Hendren (2017). For the national level results, I can implement the strategy exactly because Chetty et al. (2014) provide the national percentile transition matrix between parent and child income percentile. For state level results, I use the county-level quintile transition matrices computed by Chetty and Hendren (2017), aggregate these up to the state level, and then run a regression treating the ith quintile as the percentile $20 \times i - 10$.

For the linkage between parent income and race I use the 1980 Census 5% sample, and for the linkage between child income and race I use the 2011 and 2012 ACS (each of which is a 1% sample). I restrict the analysis to children born in the United States between 1975-1980. Notably, this specification ends up using a slightly earlier set of cohorts than those used in Chetty et al. (2014). This is partly justified by the results of Chetty et al. (2014), who find that rank-rank slopes have been very stable over recent decades, although in the future I plan to probe this and related issues further. I measure income as the total family income of the child's parents, and I construct percentiles within my sample for parent and child income (I construct percentiles separately for 2011 and 2012).[4] The Census contains the state of birth for all respondents born in the United States: I thus examine geographic heterogeneity by performing state-level analysis (I use state of birth rather than state of residence, because many children move out of state when they become adults).

I base my racial indicators on those provided by the Census. The Census distinguishes between "race" and Hispanic status. For example, a person of Mexican descent would typically show up in the Census as

[4] Following Chetty et al. (2014), I construct percentiles at the national level, rather than constructing percentiles separately for each state.

Percentile Outcome:	Child	Parent	Child	Child
Parent Percentile	—	—	0.34	0.26
Black	-22.5	-22.8	—	-18.3
Hispanic	-6.8	-12.6	—	-10.7
Asian	5.9	7.7	—	3.5
Other	-11.5	-10.9	—	-19.2
Constant	54.28	55.1	33.3	41.3

Table 1: National Level Results

First column shows regression of child income percentile on race in the 2011-2012 ACS. Second column shows regression of parent income percentile on race in the 1980 Census. Whites are the omitted category; see text for details on how racial groups are defined. Third column replicates the national level results of Chetty et al. (2014), and regresses child percentile on parent percentile. Fourth column presents estimates of the model regressing child's income percentile on parent percentile and race.

white and Hispanic.[5] Following the Census, I thus create a set of racial indicators for black, white, Asian, and other, and create a separate indicator of Hispanic status.[6] The 1980 Census only allowed respondents to list one race, while the ACS allowed for multi-racial respondents. To attempt to make these data sets comparable, I use the race of the parents in the 1980 Census rather than the listed race of the child: this allows for up to two racial groups (plus Hispanic status). For ACS respondents who list more than two races, I then recode them as the two most common races that they list (for example, a respondent who listed black, white, and Asian would be recoded as just black and white); listing more than two races is fairly uncommon. Because respondents can list multiple races, the racial indicators are not strictly colinear and thus it is theoretically possible to separately identify each indicator and a constant. However, in practice this leads to unstable results, and thus I treat the largest group, whites, as the omitted category.[7]

I begin with the national level results. These are shown below in Table 1. The first column shows the results of a regression of child income percentile on race, and the second column shows the results of a regression of parent income percentile on race; these correspond to $\hat{\delta}$ and η_{XW}, respectively, in our econometric framework.[8] As expected, blacks earn substantially less than whites, both in the sample of parents and children. There is a smaller but sizable disadvantage for Hispanics as well, and a moderate income advantage for Asians. Unlike the hispanic disadvantage, the black-white gap does not meaningfully shrink from the parent cohort to the child cohort. The third column replicates the national level results of Chetty et al. (2014), and regresses child income percentile on parent percentile, corresponding to $\hat{\beta}$ in our framework.

The fourth column of Table 1 reports the main result of the paper: estimates of the model regressing

[5] Most Hispanics show up in the Census as white Hispanics, but there are some non-whites.

[6] The Asian category includes Pacific Islanders, and the "other" category consists mainly of Native Americans.

[7] In practice, this means that respondents who respond as both white and another race are treated as simply being the other race.

[8] This is not quite accurate: since the vector of ones is contained in X, $\hat{\delta}$ and η_{XW} are actually estimates from a regression without a constant, and η_{XW} also contains a column regressing a vector of ones on race dummies. I include the constant in this table to make the estimates more readable.

child's income percentile on parent's percentile and race. The estimates reveal a sizable effect of race: on average, a black child grows up to earn 18.3 percentiles less than a white child born to a family with the same family income. There is also a sizable Hispanic disadvantage: being Hispanic lowers a child's adult income by 10.7 percentiles. The coefficient on parent income perecentile is also modestly reduced: parent income percentile takes on a coefficient of 0.26 instead of 0.34.

With these results in hand, we can answer the three questions posed in the introduction. First, how much does race matter for children's success, once we control for the effect of parent's income? The answer is that race matters a great deal: black and Hispanic children grow up to earn substantially less than white children, and most of this gap cannot be explained by parental income. Second, how much does parent's income still matter once we account for race? The results suggest that parent's income still matters a good deal, although the coefficient is reduced from 0.34 to 0.26. Third, where is the black-white gap headed? If we assume that the true model is in fact the linear model (with homogeneous slopes) that we have estimated, and we assume that the model's coefficients will remain constant over time, then we can solve for the steady-state black-white gap.[9] In the steady-state, the black-white gap will be $\frac{-\delta_{\text{black}}}{1-\beta} \approx 25$ percentiles. This suggests that we are already at the steady state: the effect of race and parent income are too strong for the black-white gap to close any further.

I next turn to the state-level analysis. Since I have not yet computed standard errors for my estimates, this analysis is much more preliminary, since some states will have very imprecise estimates due to smaller samples (for example, Vermont has only one black observation in the child sample). However, I present these results now as a proof of concept: it appears that there is regional heterogeneity that would be interesting to explore further, although in the future I will need to construct standard errors and perhaps expand my sample to more years to improve precision.

In Figure 1 I provide maps of "absolute upward mobility" by state for blacks, whites, and Hispanics. Chetty et al. (2014) define this as the expected percentile of a child born to parents at the 25th income percentile (this is their preffered measure). I use the estimated model coefficients for each state to generate predicted percentiles for non-Hispanic whites, non-Hispanic blacks, and Hispanic whites. Although preliminary, I read these maps as suggesting that once we control for race, the worst place for upward mobility becomes the Rust Belt, rather than the South (although the South is not great for mobility). This makes sense: given that there is a sizable black penalty and the South has a large black population, the poor performance of the South in standard intergenerational mobility analysis is partly driven by omitted variables bias. This is qualitatively consistent with the results of Chetty and Hendren (2017): they find that although

[9] As a simplification, this formula ignores the effect of interracial marriages. Interracial marriage will reduce the steady-state gap

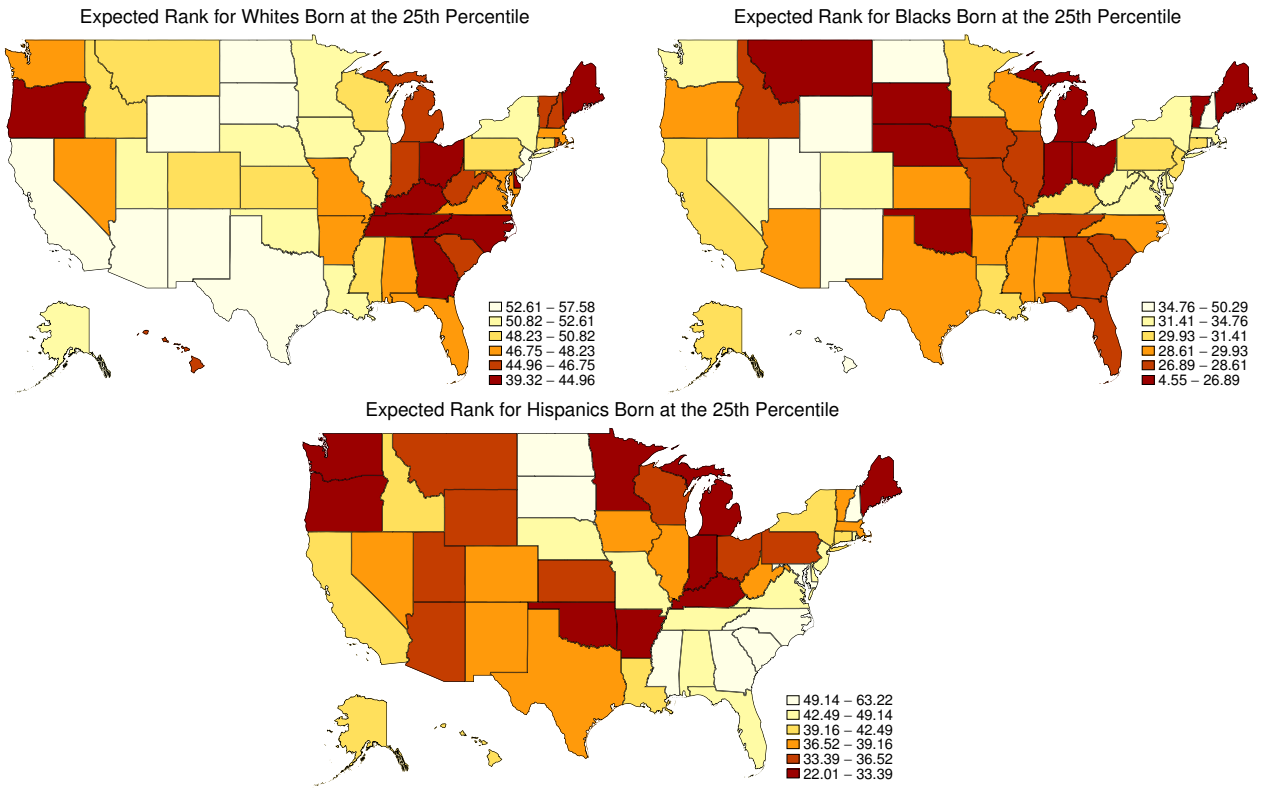

Figure 1: Absolute Upward Mobility for Blacks, Whites, and Hispanics

the non-causal estimates of upward mobility by commuting zone are highly correlated with the fraction of residents who are black, this correlation is substantially weaker when focusing on their causal estimates of upward mobility.

In Figure 2 I plot the estimated black-white mobility gap in upward mobility: since I estimate a model with homogeneous slopes, this is simply δ_{black}. Interestingly, California, Illinois, and Texas do relatively poorly, with gaps of -23, -25, and -27, respectively. By comparison Florida and New York have gaps of -20 and -19, respectively. Since these are large states their estimates are likely fairly precise. This suggests substantial heterogeneity in the black-white mobility gap, but further work is needed to better understand this heterogeneity.

Conclusion

This paper shows that it is possible to get around data limitations using econometric techniques. My preliminary results suggest that race matters a great deal for intergenerational mobility. There is a sizable black-white mobility gap ($\delta_{\text{black}} = -18.3$), as well as a Hispanic penalty ($\delta_{\text{Hispanic}} = -10.7$). Once we account for the effects of race, the importance of parent's income is reduced, but does not disappear. The

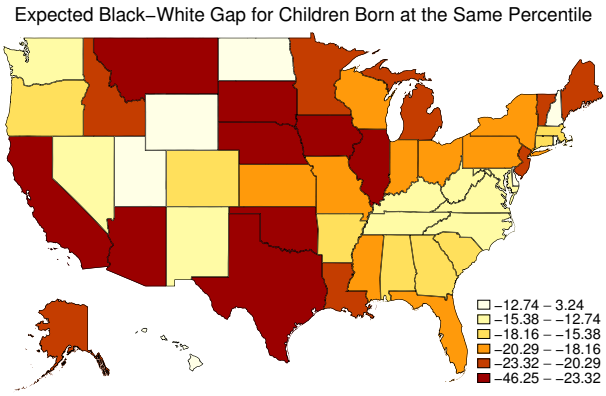

Figure 2: Black-White Mobility Gap

estimates also suggest that the black-white gap is already at the steady-state: the black-white gap is not the transient, residual effect of past discrimination, but is rather the steady-state outcome of current levels of intergenerational mobility by race. There are also interesting geographic patterns: the South performs less poorly once race is taken into account, and there also appears to be substantial heterogeneity in the black-white mobility gap.

In the future, I hope to extend the results of this paper to focus on how intergenerational mobility by race has evolved over time. Chetty et al. (2014) provide a time series for the 1971-1993 birth cohorts, using the full IRS sample for birth cohorts 1980 and later, and using the Statistics of Income (SOI) subsample (covering approximately 0.1% of tax returns) to intergenerational mobility for earlier cohorts. I plan to combine my method with their estimates to generate estimates going further back. To go back further, I have a few potential strategies in mind. One approach is to combine moments from the large samples of the Census data with moments from the smaller longitudinal data sets used by past research, in order to generate more precise estimates of intergenerational mobility by race. Davis and Mazumder (2017) use the NLS and NLSY to estimate intergenerational mobility as far back as the 1940s, suggesting it is feasible to get long time series. Going back even further, I plan to explore a method that matches synthetic parents and children using uncommon last names (essentially using last name as an instrument for parent's income/occupation). This will allow me to get accurate estimates from early Censuses (1940 and earlier) that have now been released in full with information on names. An appendix table in Chetty et al. (2014) suggests that this method works fairly well in the tax data once very common last names are dropped, because common last names are often more of a proxy for race than for income. If I am lucky, these three methods will allow me to produce a very long time series (about a century) for intergenerational mobility by race in the United States.

References

Bhattacharya, D. and B. Mazumder (2011). A nonparametric analysis of black-white differences in intergenerational income mobility in the united states. *Quantitative Economics 2*(3), 335–379.

Black, S. and P. Devereux (2011). Recent developments in intergenerational mobility. Volume 4B, Chapter 16, pp. 1487–1541. Elsevier.

Chetty, R. and N. Hendren (2017, December). The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. Working Paper 23002, National Bureau of Economic Research.

Chetty, R., N. Hendren, P. Kline, and E. Saez (2014, November). Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics 129*(4), 1553–1623.

Chetty, R., N. Hendren, P. Kline, E. Saez, and N. Turner (2014, May). Is the united states still a land of opportunity? recent trends in intergenerational mobility. *American Economic Review 104*(5), 141–147.

Davis, J. and B. Mazumder (2017, December). The decline in intergenerational mobility after 1980.

Hertz, T. (2005). *RAGS, RICHES, AND RACE: The Intergenerational Economic Mobility of Black and White Families in the United States*, pp. 165–191. Princeton University Press.

Isaacs, J. (2007, November). Economic mobility of black and white families. Brookings Report.

Kearney, M. S. (2006). Intergenerational mobility for women and minorities in the united states. *The Future of Children 16*(2), 37–53.

Mazumder, B. (2014). Blackâwhite differences in intergenerational economic mobility in the u.s. *Economic Perspectives* (Q I), 1–18.