Replication of Elizabeth O. Ananat's "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality"

Ravenna Collver, Purva Kapshikar and Casey Li

Abstract

Elizabeth O. Ananat's main goal in "The Wrong Side(s) of the Tracks: The Causal Effects of Racial Segregation on Urban Poverty and Inequality" is to identify whether racial segregation causes negative effects on economic outcomes at the city level. Ananat finds that segregation causes greater within-race inequalities and between-race inequalities (58). With only limited reference to Ananat's replication data, we rely on her original sources of data to construct a new dataset, pursue the same instrumental variables approach Ananat specifies, and only find that segregation causes an increase in Black poverty rate. These results remain significant under robustness checks that include additional controls such as education attainment and population. Finally, we consider a different causal approach by using Lin's estimator. While adjusting for some covariates produces a significant causal effect for Black and white poverty rates, adjusting for education gets rid of any significance. Overall, we fail to replicate Ananat's main result.

Introduction

This paper uses antebellum variation in the spatial arrangement of railroads at the city level to understand if segregation causes higher cross-race inequality, within-race inequality, and within-race poverty.

Segregation is an indelible part of US history. From a causal inference perspective, however, isolating the depth and breadth to which segregation limited Black Americans' outcomes is extremely difficult, since it is hard to disentangle the effect of segregation from other causes of differences in outcomes among races in America.

In order to isolate segregation, Ananat investigates pre-Civil War railroad layouts as an instrument for segregation. In the United States, most railroad tracks were laid before the turn of the 20th century, at a time in which 90% of the Black population still lived in former slave states. For states outside of the former Confederacy, segregation as a widespread social phenomena did not attain prominence until 1915 at the earliest. Afterwards, and until about 1950, deliberate government policies and collective action by white residents ensured Black and white Americans did not live, work, or play in the same neighborhoods. These discriminatory practices coincided with the Great Migration, which led some 6 million Black Americans out of the rural South and into urban regions elsewhere in the nation.

Historically, Ananat argues that urban railroads in the antebellum North were determined by factors orthogonal to segregation and any causally-related characteristic. It is true that railroad construction occurred before the Great Migration. As such, it seems chronologically improbable that railroads were laid in particular formations to encourage or discourage segregation. However, as the Great Migration occurred, railroads began to delineate racial enclaves. From an economic perspective, there is little explanation in the literature for why railroads in particular often serve as racial boundaries. Heuristically, one might expect major landmarks such as railroads to "coordinate expectations" when Black and white parties often conflict and share only limited communication (Thomas C. Schelling 1963). In other words, railroad layouts, although initially distributed as good as randomly, help facilitate segregation. For example, in cities where railroads do not divide the area into clearly parceled neighborhoods, it may be more difficult for segregation to take root, since residents, police, real estate agents, and banks may not agree on what is a Black neighborhood and what is a white neighborhood. It is precisely this historical relationship Ananat exploits in her instrumental variables approach.

Data

Ananat collects data for 121 Metropolitan Statistical Areas (MSAs) to use in this paper. According to Ananat, her data comes from three main sources:

1. The Census (microdata from IPUMS and other aggregated Census reports) for information on race, educational attainment, and income;
2. Cutler and Glaeser (1997) and Cutler, Glaeser, and Vigdor (1999) for information on segregation;
3. The Harvard Map Library paired with ArcGIS for geographic data on city layouts. This geographic data was used to calculate her Railroad Division Index (RDI), which quantifies the extent to which railroads divide cities into discrete neighborhoods.

We used geographic data directly from her publication data due to the inaccessibility of digital records from the Harvard Map Library and unfamiliarity with ArcGIS. We got the Cutler, Glaeser and Vigdor, or CGV, data directly from their archived site.[1] The Census data we acquired comes from historical Census reports or from IPUMS microdata. We found many covariates used in falsification and robustness checks in the historical Census reports, while population and income information for the main results was found using IPUMS microdata.

Summary statistics

In order to compare Ananat's data with what we were able to collect, we evaluated some summary statistics for several outcome and control variables for the 121 cities she considered in her sample. We have included these tables and figures in our appendix.

Table 1A contains summary statistics for the 1990 distributional characteristics of segregation and RDI for those cities included in Ananat's sample. In particular, it is interesting to note that our in-sample cities are, on average, not very Black: the sample cities average 6.14% Black versus the US 1990 mean of about 12%.[2] Additionally, the in-sample cities are heavily skewed towards being more divided by railroads than not: The distance between the minimum

[1] http://web.archive.org/web/20090605144330/http://trinity.aas.duke.edu/~jvigdor/segregation/index.html
[2] https://www.census.gov/prod/cen1990/wepeople/we-1.pdf

and first quartile of RDI was .4, whereas the distance between the third quartile and maximum of RDI was only .16.

Table 1B contains mean characteristics of cities in and out of Ananat's 121-city sample, with columns for our replication beside Ananat's original calculations. The in-sample mean characteristics are nearly identical, whereas the out-of-sample ones are different. (Here, "in sample" refers to cities whose railroad maps were stored in the Harvard Map Library and thus were included in the main results. "Out-of-sample" refers to those cities that were included in the CGV data but not in Ananat's selected 121 cities.) Ananat performs *t*-tests to show that the in-sample and out-of-sample cities are not significantly different from one another, but we do not draw the same conclusion. For example, we find that the cities Ananat considers are significantly less Black, population-wise, than the cities not in sample — all the *p*-values for *t*-tests performed on the differences in mean percentage of population that is Black between the two groups are statistically significant. Intuitively speaking, we interpret this result to indicate that whether or not a city was sampled was not independent of racial demographics.

The discrepancies likely arose from differences between what Ananat and we considered "out-of-sample." We considered all 352 observations from the original CGV dataset as our total population of cities, of which 121 cities were considered in-sample, and the remainder out of sample. However, upon inspection of Ananat's data, we find she begins with 367 total cities. We have been unable to identify from where this underlying difference in data arises.

Figure 1C, on the relationship between RDI and segregation for sampled cities, is an attempt to replicate Figure 1D (Ananat's original graph). These two figures are, as far as we can tell, identical, and corroborate Ananat's claims that RDI is an adequately strong instrument for segregation. Tables 1E and 1F are summary statistics for other variables we use in our analysis, as outcomes, covariates and the instrument.

Discussion of data issues

While we strove to acquire data from the same sources as Ananat, we ran into many issues in doing so. First and foremost, we were unsure which year Ananat's outcome data came from. Ananat does not explicitly state what year these data are from, so we used data from 1990 in accordance to the year the CGV segregation data was gathered. Additionally, Ananat's falsification checks specifically controlled for MSA characteristics from 1990 and 1920, which seemed like additional evidence that we should use 1990 data.

Secondly, the sources of Ananat's raw data is unclear: Ananat says that "aggregate city income distributions by race [are measured] using poverty rates from published census reports and measures of inequality generated from public-use microdata" (48). We hand-collected poverty rates by race and MSA from reports published by the Census Bureau on data from the 1990 Census[3] and calculated income percentiles and Gini indices by race and MSA from 1990 IPUMS microdata. At this point, our hand-collected poverty data and calculated income percentile and Gini index data did not match with Ananat's ICPSR replication data, but without more detail on which specific Census publications and data sources Ananat used, we deemed the 1990 Census report and 1990 IPUMS data sufficient.

Thirdly, Ananat indicates that she did some geographic crosswalking[4] to control for the effects of urban growth, which caused what were once distinct cities with distinct railroad systems to merge into single MSAs. Ananat writes that she uses "MSA-level data for the 64 cities that have remained independent MSAs, … uses county-level data for MSAs in which city centers are each in a separate county, … and [assigns] the characteristics of the politically defined city itself for MSAs that share a single county with another city" (42). However, with our lack of clarity on which data sources Ananat initially referred to, we were unable to crosswalk 1990 MSAs to their historical cities, since the 1990 IPUMS data only records MSAs and not cities.

A major problem with the resulting data is that we lack information on 17 of the cities that Ananat has data for. The 1990 IPUMS microdata only recorded survey responses from 104 unique MSAs, but Ananat's ICPSR data records observations for 121 MSAs. These 17 cities are missing because the Census did not publish data on MSAs with populations below 250,000, and IPUMS had data on even fewer MSAs with small populations. We are not sure how Ananat got data on these smaller MSAs. Publicly-available data omits the MSA variable for Census respondents who live in small MSAs to protect their privacy.

Not only do these 17 missing cities form a sizeable portion of data, but there may be systematic differences about these MSAs (e.g., they only gained enough population after 1990 to

[3] https://www2.census.gov/library/publications/decennial/1990/cph-l/cph-l-107.pdf

[4] https://www.nhgis.org/user-resources/geographic-crosswalks

become classified as MSAs) that would have significantly changed the overall regression results. Additionally, IPUMS data is skewed towards those living in city centers — respondents living in the outskirts of MSAs that overlap with another often cannot be conclusively attributed to one MSA or the other, and, as such, are excluded from both MSA's aggregates. This weakness in the data may be especially relevant if there are systematic differences between inhabitants of the center of a MSA and inhabitants who live at the edges of a MSA. Given that this paper studies the spatial distribution of people of different races within an MSA, this weakness may be difficult to overlook.

Assumptions

This paper examines the causal effect of racial segregation on city-level economic outcomes. Ananat's approach is an approximation of an ideal randomized experiment of a perfectly segregated city and perfectly integrated city that are otherwise identical. Residents would be assigned to these cities randomly from initial distributions. In this experiment, the relationship between segregation and offspring's income distribution indicates the treatment effect of segregation. The selection effect of segregation could then be determined by measuring demand for cities after residents are allowed to move.

To approximate this ideal setup, Ananat proposes a two-stage least squares approach using railroad track configuration as an instrument for segregation. Segregation is measured by Cutler, Glaeser, and Vigdor's 1990 dissimilarity index. It ranges from 0, perfect integration, to 1, perfect segregation. Railroad track configuration is measured by RDI, which measures to what extent a MSA is spatially divided by railroad tracks into distinct neighborhoods. It is related to the Herfindahl index, and ranges from 0, when one large neighborhood forms the entirety of the MSA, to 1, when infinitely-many infinitely-small neighborhoods form the MSA. In all models that she runs, she controls for the length of railroad track per square kilometer. The two-stage least squares approach relies on three main assumptions, according to her paper:

1. RDI induces meaningful variation in the degree of racial segregation, given the control. In order for this assumption to hold, the coefficient of RDI in the following regression should be significant:

$$Segregation = \beta_0 + \beta_1 \cdot RDI + \beta_2 \cdot Density\ of\ track$$

In this regression, *Segregation* refers to the 1990 dissimilarity index. *Density of track* is a covariate controlling for the length of railroad track per square kilometer in each MSA.

2. RDI affects city outcomes only through racial segregation, i.e., there is no direct relationship between track configuration and city outcomes, given the control variables. Ananat only uses the density of track as a control, so for this assumption to be valid, RDI should not be correlated with any city outcome given track density. Ananat argues that before the Great Migration (before racial segregation could have significant direct effects of human capital), city characteristics were not affected by segregation, so there should be no correlation between RDI and 1910 city characteristics. Therefore, the coefficient of RDI in the following regression(s) should not be significant:

$$Y = \beta_0 + \beta_1 \cdot RDI + \beta_2 \cdot Density\ of\ track$$

In this regression, the outcome Y is one of the following 1910 city characteristics: physical area, population, ethnic dissimilarity index, ethnic isolation index, percent Black, or streetcars per capita. If RDI is found to be correlated with the 1910 city characteristics, one option would be to include that characteristic as an additional control.

Ananat gives three additional reasons why this assumption should be valid in the form of refuting three counter arguments. The first is that RDI could affect city outcomes through regional geographic variation. To test whether this is true, Ananat replicates her main results while controlling for the Census region. If the results are the same, then the assumption is still valid. The second counter argument is that RDI could reflect the value of land in an area, which affects city outcomes since low property values could lead to segregation by income, which would appear as segregation by race. While she says this is historically implausible, she also runs a regression of income segregation in 1990 on RDI. If RDI does not have a significant effect on income segregation, then this assumption is valid. Thirdly, there could be some unknown channel through which RDI affects segregation. To test this, Ananat controls for various other covariates when running her main results. If the results don’t change, then this assumption is valid.

3. Characteristics of residents and RDI must not be correlated with city characteristics, given the control variable, to approximate the ideal experiment with quasi-random city assignment. For this assumption to be valid with only track density as a control, the coefficient of RDI in the following regression(s) should not be significant:

$$Y = \beta_0 + \beta_1 \cdot RDI + \beta_2 \cdot \textit{Density of track}$$

In this regression, the outcome Y is one of the following 1920 human capital characteristics: percent Black, percent literate, labor force participation rate, or share of employment in trade, manufacturing, and railroads.

If these three assumptions hold, railroad configuration is not correlated with underlying city or population characteristics given the density of track, so RDI can be used as an instrument in two-stage least squares with the one control.

Appraisal of the three stated assumptions

1. Controlling for the density of track in historical city centers, neighborhood RDI is a strong predictor of the degree of racial segregation, as seen in Table 2A of the appendix. We replicate this first stage in our regression results and find a statistically significant coefficient and a F-statistic of 15.07. Following the general rule used in applied economics that an F-statistic greater than 10 is a strong indicator for the relevance condition to hold, we conclude that Ananat's assumption holds. We discuss these regression results in greater detail in the following section.

2. From historical accounts and more recent analysis, it does not seem that 19th century railroad configuration was driven by social or economic motivations, but by orientation of nearby locations.[5] Furthermore, in 1910, after the end of major railroad construction and before the Great Migration, RDI was unrelated to the selected city characteristics after controlling for density of railroad track. We report the coefficients and significance levels of the regressions of city characteristics on RDI in Table 2B of the appendix. RDI did not have a significant effect on any outcome that we tested, since all of the coefficients in the regressions we ran were not significantly different from zero. This includes the coefficient of the regression of income segregation in 1990, which shows that RDI did not affect city outcomes through land value. However, the number of cities in the sample is limited to what the CGV dataset provides, so it may be small enough for us to distrust the results that were deemed significant.

She additionally replicated the main results with indicator variables for Census region along with several other covariates: while the standard errors of the estimates increased, the

[5] Ananat (2011); Atack and Passel (1994); Wellington (1911). We discuss this further in the appendix, under the section "Assumptions".

results were almost the same. Based on Ananat's results for these regressions, RDI only affects outcomes in cities with significant Black inflows, so it has no direct relationship with current city characteristics.

3. From Ananat's results, in 1920, just after the first wave of the Great Migration, RDI was unrelated to the selected human capital characteristics of cities. We replicate this in our robustness checks later.

Results

Ananat's main results consist of ordinary least squares regressions (OLS) and two-stage least squares regressions (2SLS) on eight outcome variables: Gini indices and poverty rates for Black and white populations within a MSA and the following ratios of the white and Black income distributions within a MSA: the 90th percentile of white incomes and 90th percentile of Black incomes; the 10th percentile of white incomes and 90th percentile of Black incomes; the 90th percentile of white incomes and 10th percentile of Black incomes; and the 90th percentile of Black incomes and 10th percentile of white incomes. We take the log of every outcome variable with the exception of poverty rates, as Ananat does.

Ordinary least squares

In Ananat's ordinary least squares regressions, she regresses one of eight outcome variables listed above directly on segregation:

$$Y \ = \beta_0^{OLS} + \beta_1^{OLS} \cdot Segregation$$

These OLS results are secondary to results from the 2SLS regression. The exact coefficient estimates and robust standard errors are listed in Table 3 of the Appendix alongside Ananat's results.

Intuitively speaking, our OLS regressions indicate that a one-standard-deviation (14 point) increase in segregation causes a 1.3 percentage point decrease in white poverty rates (significant) and a 3.8 percentage point increase in Black poverty rates (significant). Analogously, a one-standard-deviation increase in segregation causes a 1.7 percentage point increase in white Gini index (insignificant) and a 6.2 percentage point decrease in Black Gini index (significant).

First stage

In the first stage of the 2SLS, our model is as detailed by Ananat as:

$$Segregation = \alpha_0 + \alpha_1 \cdot RDI + \alpha_2 \cdot Density\ of\ track + \varepsilon$$

Our OLS estimates for α_1 and α_2 are 0.3755 and 17.9535, with robust standard errors of 0.084 and 9.061, and are statistically significant and insignificant, respectively. We estimate α_0 to be 0.2816.

Intuitively speaking, the results of this regression indicate that even a perfectly undivided MSA with zero railroad track is somewhat segregated. This conclusion suggests that segregation occurs even without the aid of railroad tracks and their division of a MSA into smaller neighborhoods. Additionally, altering the placement of railroad tracks within a MSA without changing the overall density of track, so that a perfectly undivided MSA becomes a perfectly divided MSA, is associated with an increase in segregation by 0.3755.

We report a F-statistic of 15.07, but a somewhat small R-squared value of 0.23. Although a F-statistic greater than 10 does not indicate a weak instrument, the small R-squared value may indicate additional issues. We cannot ascribe these results to differences in data processing: we have used the CGV data exactly as Ananat has indicated in the original paper, and we have directly used Ananat's own measurements of RDI and density of railroad track in each MSA due to inability to access historical railroad maps in the Harvard Library.

Second stage

The model for the second stage of the 2SLS is given below. Note that $\widehat{Segregation}$ refers to the fitted values of $Segregation$ from the first stage.[6]

$$Y = \beta_0 + \beta_1 \cdot \widehat{Segregation} + \beta_2 \cdot Density\ of\ track\ ^{7}$$

Intuitively speaking, our 2SLS regressions indicate that a one-standard-deviation (14 point) increase in segregation causes a 1.4 percentage point decrease in white poverty rates

[6] Note that in our code, we do not run two separate regressions, but instead use the Python module statsmodels' IV2SLS function. The two stages are written out explicitly here for clarity.

[7] Note that "Density of track" here is shorthand for "within-MSA mean length of track per square kilometer." It is referred to elsewhere in this writeup as "mean length of track per square kilometer." This data was hand calculated by Ananat using records from the Harvard Map Library. As such, we directly use her estimates in our empirical work.

(insignificant) and a 5.3 percentage point increase in Black poverty rates (significant). Analogously, a one-standard-deviation increase in segregation causes a 0.28 percentage point increase in white Gini index (insignificant) and a 1.8 percentage point decrease in Black Gini index (insignificant).

Unlike in Ananat's original work, wherein all of her 2SLS estimates are greater in magnitude than her OLS estimates, many of our estimates from the 2SLS regressions are smaller in magnitude and insignificant compared to our OLS estimates. Notably, only one of our 2SLS-estimated causal effects is statistically significant (Black poverty rate) whereas seven out of Ananat's eight are. Additionally, although the signs of our OLS and 2SLS estimates align with those of Ananat when both are significant, the difference in magnitude between our estimates and Ananat's estimates can be quite sizable (see Table 3A in Appendix). We ascribe these differences to several data issues, which are explored in depth in the earlier "Data" section of this paper, under the heading "Discussion of data issues." Overall, we do not replicate Ananat's original results.

Robustness Checks

Ananat's chosen robustness check is to repeat her main result while controlling for various 1920 and 1990 city characteristics. Doing so allows her to determine whether RDI affects city outcomes directly or only through these other city characteristics. If the results change when she includes these other characteristics, then the proposed causal effect of segregation on city outcomes breaks down, since the other characteristics could drive the causal relationship rather than segregation. Testing for selection on unobservables using selection on observables is a fairly common practice. Heuristically speaking, if unconfoundedness requires more observable variables to hold, then unconfoundedness may require more unobservable variables to hold as well. Ananat finds that her results do not change when including these additional covariates: the estimated causal relationship does not differ significantly from the relationship when she does not include them. She concludes from these findings that her identified causal relationship is due to segregation.

Ananat includes 1990 and 1920 population, percent of the population in 1990 and 1920 that is Black, education levels by race in 1990 and overall literacy rates in 1920, share of the total workforce in manufacturing in 1990 and 1920, labor force participation by race in 1990 and overall in 1920, the number of local governments formed by 1962, and a propensity score (more explanation below) as covariates for each MSA. The 1990 “Education” variable includes controls for the percentages of high school dropouts, high school graduates, college graduates, and people who have completed some college, broken down by race. For clarity, we refer to the controls as “1990 city characteristics" and “1920 city characteristics.” The regression results in Table 3 report the coefficient and standard error associated with the 2SLS regression of an output, given by the column label, on segregation (which is in turn instrumented on RDI) and any control variable as indicated by the row label.

Ananat acquired data for these checks from Census publications and IPUMS microdata, with the exception of the number of local governments, which came from the CGV data. We acquired our data from the same sources, using CGV data for 1990 manufacturing share and the number of governments, IPUMS microdata for 1920 labor force participation and manufacturing share, and Census publications for the rest. As noted in the Data section, our data contains 17 fewer MSAs than Ananat’s.

Ananat’s propensity score was generated by estimating the probability of a city having an above-median RDI given the 1920 city characteristics and the distance of that city from the South. We used a logistic regression model with those variables to find the propensity score. (We assume, given the era this paper was written, that Ananat also chose to use a logistic regression as opposed to a nonparametric technique).

The coefficient estimates from Ananat’s results after controlling for 1990 characteristics were not significantly different from those estimated for her main results, although many of these estimates and their standard errors did increase slightly. This led her to conclude that the instrument of RDI impacts poverty and inequality, which were represented by the outcome variables of poverty rates and Gini index by race, respectively, solely through segregation. The coefficient estimates after controlling for 1920 characteristics are also not significantly different from her main results and, just as in her main results, are all statistically significant.

In our replication, we found similar, but not exact, results. Specifically, we found that controlling for population, percent Black, and share in manufacturing in a MSA in both 1990 and 1920 did not significantly affect the results found in the main results. However, controlling for education levels and labor force participation in 1920 and 1990 did affect the main results. In 1920, this difference between results and replication may be due to omissions of MSAs due to missing data (121 MSAs in Ananat's regression compared to only 80 and 49 MSAs for literacy and labor force participation, respectively, for ours). In 1960, however, there was comparatively less difference in data: We were only missing a single MSA. These results suggest that education and labor force participation, at least in 1990, could explain the association between segregation and city outcomes, rather than segregation directly causing this relationship.

The results for the Black poverty rate outcome are always significant, no matter which additional covariate we control for. These results indicate that if there is a latent variable causing segregation to be associated with a higher Black poverty rate, it is not one of the variables we have controlled for. This suggests that segregation could cause higher Black poverty rates, even if it does not cause the observed relationship for the other outcomes.

Reanalysis

In short, our prior work fails to replicate Ananat's main result. Among both model specifications (OLS and 2SLS) and with all eight outcomes, only our OLS model of white and Black poverty rates replicates Ananat's results in the correct direction and at a 0.05 level of statistical significance.

In general, we ascribe this failure to data procurement, which we have discussed in detail in an earlier section titled "Data," under the subheading "Discussion of data issues".

Reanalysis of main results with 17 missing cities

Using those MSAs we were able to find all the data for (n=104) and those that we could not find in IPUMS (n=17), we ran a series of two-sample t-test (with unequal variance) on each outcome. We hoped to determine whether there were any significant differences in these two groups which may have affected our main results, since our coefficient estimates calculated from data on only 104 MSAs were different from Ananat's. We computed the difference in means and significance for the eight outcome variables, endogenous variables, and instrument. In order to avoid the appearance of systematic differences between these two groups of cities that are actually due to differences in data procurement, we have compared these two groups of cities as they appear in Ananat's ICPSR data. The results are included in Table 5B in our appendix.

The difference in means of these two groups is only significant for the length of railroad track per square kilometer. The mean length of railroad track for the included cities was slightly larger than that of the missing cities. This difference, intuitively, may be accounted for by the historical role of railroads in population growth. We think that the MSAs that were not in our dataset were excluded from IPUMS indirectly due to their lack of population. Cities that did have large enough populations to be included in IPUMS likely did so due to larger railroad networks.

In order to determine whether the absence of these 17 MSAs drove our inability to replicate Ananat's results, we ran two additional analyses. In Table 5C of our appendix, we include four sets of results from running ordinary least squares and two-stage least squares regressions. The equations for these regressions are included in the earlier section for our main results.

Of the four sets of results, two of these, the columns named "Ananat" and "Group 3," have been copied from our main result. They are Ananat's results from her paper and our main result replication using our collected data of 104 MSAs. The other two are the additional analyses: the column named "Ananat - 17" contains results from using Ananat's data for only those 104 MSAs that we were able to find data for, meaning that we removed those 17 cities (listed above) before running the regressions on her data. The column name "Group 3 + 17" includes the results from running the regressions using our data for the 104 MSAs we were able to find all the data for, in addition to the 17 missing MSAs (by using Ananat's data for those directly), to, ideally, match her dataset of all 121 MSAs. We recognize that in the "Group 3 + 17" data there could be issues, as our data has not always seemed to match Ananat's in scale.[8] However, the significance of the results is almost consistent across Ananat's and our data, even with these new analyses. The results from running these regressions is included in Table 5C of our appendix.

We see that trying to account for the 17 missing cities, even in two different ways, has only slightly affected our results. Only three coefficients have changed in significance: 1) The coefficient of the cross-race income ratio for the 90th percentile becomes significant in the OLS results for our new data ("Group 3 + 17"), which no longer matches our previous or Ananat's results. 2) The coefficient of the income ratio for the 90th percentile for the Black population to

[8] To clarify, the raw scale of data from Ananat's ICPSR data and our data seemed somewhat off, especially with regard to variables sourced from IPUMS. For instance, our weighted within-race, within-MSA mean annual incomes were consistently lower than Ananat's measures. That said, the ultimate outcome variables are transformed so that these across-MSA differences in measurement are attenuated: For example, incomes percentiles for races within a MSA are divided to get ratios before being logged, which hopefully attenuates the difference in scales between Ananat's dataset and our dataset and facilitates the neat transplant of Ananat's 17 cities into our dataset.

the 10th percentile for the white population loses significance in the OLS results with our new data ("Group 3 + 17"), which does match Ananat's results. 3) From our 2SLS regression, this same coefficient loses significance when only considering the subset of 104 cities from Ananat's original data ("Ananat - 17"), which does not match Ananat's main results but does match the our results from both our original and new data. Because these effects are slight and not systematic across all variables, we believe that the results from these re-analyses indicate that we, on the whole, still fail to recreate Ananat's effects of segregation on these outcomes.

Dichotomizing to use Lin's estimator

We reanalyze Ananat's results another way by abandoning the instrumental variables framework and employing Lin's estimator. In particular, we choose Lin's estimator with Eicker-Huber-White robust standard errors because we have many covariates for which we can adjust. This estimator has a smaller bias than Fisher's ANCOVA estimator with a conservative standard error estimate. In this section, we perform analyses with both our collected data (n=104) to test the robustness of our main result replication and with Ananat's ICPSR data (n=121) to test the robustness of her results with regard to method.

Because our instrument (RDI) and treatment variable (segregation) are both continuous, in order to be able to analyze the data using methods from the course we had to first dichotomize both variables. To do so, we split the instrument and the treatment at the median, since Ananat used the median cutoff in her paper to generate propensity scores. With our dichotomized data, we ran OLS on the following equation, where Y is one of the city outcomes, Z is the dissimilarity index of segregation, and X is various combinations of centered covariates:

$$Y_i = \beta_0 + \beta_z Z_i + \beta_x{}^T X_i{}^T + \beta_{zx}{}^T Z_i X_i{}^T + \varepsilon_i$$

The estimated coefficient of Z is equivalent to Lin's estimator for the causal effect of segregation on city outcomes. The standard error of the estimate is the Eicker-Huber-White robust standard error for the OLS estimate, found in R using the `hccm()` function. The results for various choices of X, our covariates, are included in Table 5D in our appendix.

Different choices of covariates produced varying results. For the most part, the regressions failed to produce a significant causal effect. This backs up our main replication result: Segregation does not appear to have a direct causal effect on any city outcome except perhaps Black poverty rates.

However, a few results stand out. Firstly, adjusting for the length of track per square kilometer only produces a significant effect for white poverty rates and Black Gini indices. This result intuitively contradicts Ananat's results and corroborates our replication: In Ananat's original OLS and 2SLS estimates, her only control variable is the length of track per square kilometer. In our replication of Ananat's 2SLS, we found no significant effect on any outcome variable but Black poverty rates. In our application of Lin's estimator, we only find two significant effects of segregation on city outcomes.

Secondly, adjusting for the propensity score, for the 1990 population, and for the 1990 and 1920 populations resulted in a significant negative effect of segregation on white poverty rates. These results match Ananat's results, though our two-stage least squares replication failed to recreate the effect. However, adjusting for the 1990 education characteristics did not produce the significant negative effect. This suggests that education could be driving the relationship between segregation and white poverty: perhaps cities with a more educated white population are more segregated, and the higher level of education results in lower poverty rates for white residents.

Other outcomes also saw varying degrees of significance. For example, the population and percent of Black residents in 1990 and 1920 were not enough to explain away the relationship between segregation and Black poverty rates. However, adding in education characteristics resulted in an insignificant effect, as for white poverty rates. In fact, when controlling for education, segregation was not found to have a significant effect on any city outcome. This suggests that education is a significant confounder for the relationship between segregation and city outcomes.

Overall, using Lin's estimator with the dichotomized above-median segregation indicator failed to produce consistent results that matched Ananat's. No effect was significant when adjusting for the education levels. This result fits with our overall finding that segregation does not have a causal effect on city outcomes.

Moreover, we acknowledge that any differences between our Lin's estimate of causal effects and Ananat's original 2SLS estimates may be driven by differences in data procurement. As such, we also used the same approach of dichotomization and Lin's estimator on Anant's ICPSR data, and found that adjusting for 1) education levels in 1990, 2) education levels in 1990 and population in 1990 and 1920, and 3) education levels in 1990, population in 1990 and 1920, and percent Black in 1990 and 1920 all yield insignificant causal effects. In contrast, Ananat, in her robustness checks, controls for these variables and still finds a significant causal effect of segregation. In short, simply switching methods—from two-stage least squares to Lin's estimator (both with the same sets of controls) causes the causal effect of segregation to disappear.

Conclusion

While Ananat found a causal effect of racial segregation on city outcomes using two-stage least squares, we were unable to replicate this result. We ascribe this to issues in data collection, since the sources Ananat mentions do not contain all of the information needed to replicate her main results, robustness checks, or falsification checks. The closest we came to replicating her results was that we found, as Ananat did, that segregation increases Black poverty rates. However, when appraising Ananat's conclusions, we found that controlling for education in the two-stage least squares results in an insignificant effect for poverty rates and Gini indices. We also found that Lin's estimator fails to produce a significant effect on any outcome when controlling for education. Therefore, we cannot conclude that there is a causal effect of segregation on economic outcomes.

Appendix

Summary statistics

I. Measurements of segregation

Racial isolation and racial dissimilarity are instruments of segregation from Cutler, Glaeser, and Vigdor (CGV) and RDI is a constructed measurement from Ananat herself. These measurements all have a theoretical range, where 0 is perfect integration for racial isolation and racial dissimilarity (i.e. Black populations are randomly distributed throughout a city's census tracts), and perfectly undivided by railroads for RDI (i.e., a city is one large neighborhood), and 1 is perfect segregation for racial isolation and racial dissimilarity, and 1 is perfect division (i.e., a city is divided into infinite neighborhoods of zero area).

A. Table 1A: Summary statistics for in-sample segregation measures (1990)

	Isolation	Dissimilarity	% Black	Population	Pop. Black	Area	RDI
Mean	0.212	0.568	0.061	590189	54294	1825.886	0.723
Std. dev.	0.189	0.136	0.052	1062319	154547	2882.564	0.141
Minimum	0.006	0.329	0.005	70683	1044	89.600	0.238
Q1	0.036	0.452	0.018	158983	3236	580.600	0.638
Median	0.185	0.574	0.049	273064	13770	1031.100	0.742
Q3	0.349	0.673	0.093	480628	38640	1926.100	0.839
Maximum	0.763	0.873	0.232	8863164	990406	27269.900	0.987

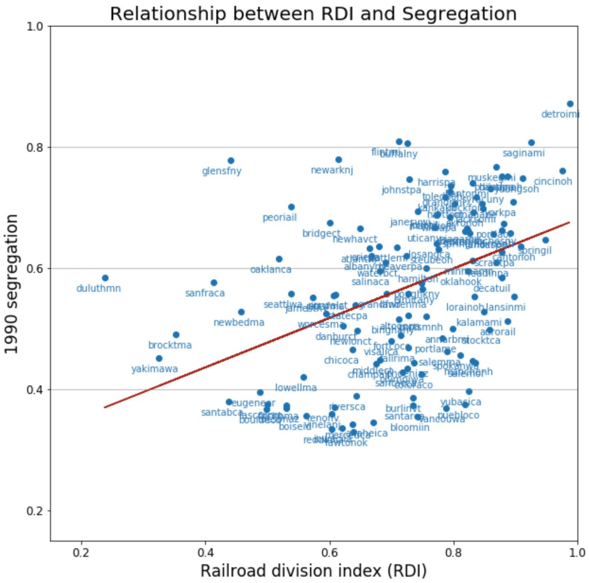
B. Table 1B: Mean characteristics of cities in and out of sample

The "Difference in Means" column gives the difference in means between cities in and out of the sample. The values provided in parentheses are the standard deviation of the city characteristics. Bolded cells indicate that the sample means differed significantly for those characteristics at the 0.05 level. Note that the difference in means is not comparing our data with Ananat's, but instead comparing the means of the two samples of cities across her and our data separately.

CGV variable		Not in sample		In sample		Difference in means	
		Group 3	Ananat	Group 3	Ananat	Group 3	Ananat
Isolation index	1890	0.059 (0.003)	0.049 (0.007)	0.053 (0.003)	0.053 (0.008)	0.007	-0.004
	1940 (tract)	0.423 (0.013)	0.355 (0.053)	0.318 (0.018)	0.318 (0.043)	0.106	0.037
	1940 (ward)	0.223 (0.009)	0.234 (0.034)	0.198 (0.013)	0.198 (0.023)	0.025	0.036
	1970	0.440 (0.014)	0.343 (0.034)	0.359 (0.018)	0.365 (0.023)	**0.081**	-0.022
	1990	0.282 (0.012)	0.229 (0.002)	0.212 (0.017)	0.214 (0.017)	**0.070**	0.015
Dissimilarity index	1890	0.310 (0.008)	0.385 (0.032)	0.383 (0.009)	0.383 (0.024)	**-0.073**	0.002
	1940 (tract)	0.704 (0.008)	0.736 (0.029)	0.742 (0.008)	0.742 (0.019)	-0.038	-0.006
	1940 (ward)	0.502 (0.010)	0.570 (0.032)	0.579 (0.012)	0.570 (0.022)	**-0.068**	0.000
	1970	0.720 (0.008)	0.744 (0.015)	0.738 (0.010)	0.740 (0.012)	-0.018	0.004
	1990	0.553 (0.008)	0.574 (0.016)	0.568 (0.012)	0.569 (0.012)	-0.015	0.005
Percent Black	1890	0.209 (0.013)	0.030 (0.005)	0.027 (0.002)	0.027 (0.003)	**0.182**	0.004
	1940	0.171 (0.009)	0.058 (0.007)	0.041 (0.003)	0.041 (0.005)	**0.130**	**0.018**
	1970	0.133 (0.007)	0.056 (0.006)	0.062 (0.004)	0.062 (0.005)	**0.072**	-0.006
	1990	0.131 (0.007)	0.067 (0.006)	0.061 (0.005)	0.061 (0.005)	**0.070**	0.005

C. Figure 1C: Relationship between RDI and segregation, replicated "1990 segregation" on the vertical axis refers to dissimilarity index as calculated in 1990 from the CGV data.

D. Figure 1D: Relationship between RDI and segregation, Ananat's original



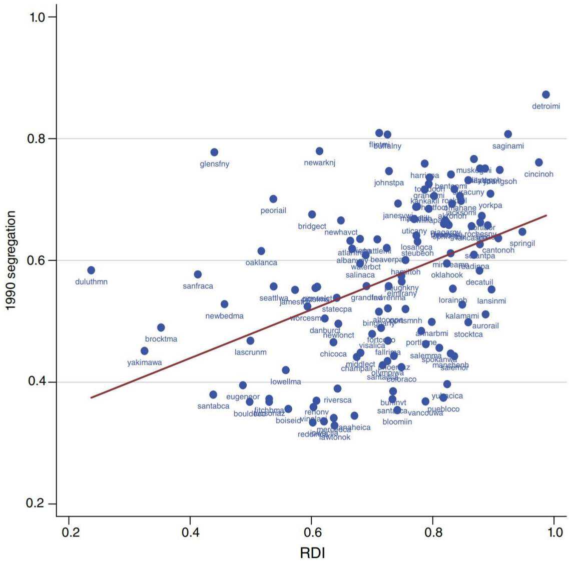


FIGURE 3. FULL SAMPLE RELATIONSHIP BETWEEN RDI AND SEGREGATION

II. Additional summary statistics

A. Table 1E: Poverty rate (1990 Census)

	Poverty rate	
	White population	Black population
Mean	0.094	0.287
Standard deviation	0.035	0.092
Minimum	0.034	0.064
Q1	0.069	0.213
Median	0.092	0.291
Q3	0.111	0.354
Maximum	0.262	0.532

B. Table 1F: Railroad index, length and nearest former slave state (Ananat's original data)

	RDI	Density of track	Closeness
Mean	0.723	0.0009	-362.434
Standard deviation	0.141	0.0013	331.822
Minimum	0.238	0.0002	-1163.790
Q1	0.638	0.0004	-398.540
Median	0.742	0.0007	-241.740
Q3	0.830	0.0010	-150.970
Maximum	0.987	0.0132	-13.000

Assumptions

A. Table 2A: Assumption 1

The following table includes the coefficients and robust standard errors (in parentheses) for the predictor of RDI in:

$$Segregation = \beta_0 + \beta_1 \cdot RDI + \beta_2 \cdot Density\ of\ track$$

Bolded cells indicate coefficients significantly different from zero at the 0.05 level.

Outcome	Group 3	Ananat
1990 metropolitan dissimilarity index	**0.375** **(0.084)**	**0.357** **(0.088)**

B. Table 2B: Assumption 2

The following table includes the coefficients and robust standard errors (in parentheses) for the predictor of RDI for the city characteristics (Y) listed in the first column:

$$Y = \beta_0 + \beta_1 \cdot RDI + \beta_2 \cdot Density\ of\ track$$

Outcome	Group 3	Ananat
Physical area (square miles/1000) 1910	-4.479 (24.712) n = 43	-3.993 (11.986) n = 58
Population (1000s) 1910	68.199* (454.542) n = 46	0.666 (1.36) n = 121
Ethnic dissimilarity index 1910	-0.225 (0.222) n = 29	0.076 (0.185) n = 49
Ethnic isolation index 1910	-0.140 (0.148) n = 29	0.027 (0.070) n = 49
Percent Black 1910	-47.1* (53619) n = 46	-0.0006 (0.0100) n = 121
Street cars per capita (1000s) 1915	-0.132 (0.246) n = 13	−0.132 (0.183) n = 13
Income segregation 1990	0.032 (0.058) n = 69	0.032 (0.032) n = 69

*Although the values are quite different, our results match Ananat's in significance.

C. Historical background for 19th century railroad configuration

Ananat states that 19th century railroad configuration is driven by orientation of locations. She cites historical accounts and more recent analysis to support this claim; for instance, these accounts suggest that the configuration of tracks was optimized based on both orientation of nearby destinations and variation in ground slope (Wellington 1911; Atack 1994). She compares two cities, Binghamton, New York and York, Pennsylvania, as having developed around the same time, with similar geographies and even roughly equal total kilometers of track. However, the configurations differ in these two cities due to the configurations of more distant hills. Ananat also provides further detail in her appendix, stating that the three main drivers of United States railroad track configuration and placement were slope, competitive strategy and national security. She explains that historical accounts do not mention railroad track configurations being used or having effects as a "social barrier" at the time they were being laid. Given these explanations, in addition to the results in the table above, we believe that Ananat's second assumption holds.

Results

A. Table 3A: Main results

The following table reports our replication of Ananat's main results alongside her results. The original table can be found on page 53 of Ananat's original publication in *AEJ*. Bolded cells indicate coefficients significantly different from zero at the 0.05 level.

Outcome	Race	OLS		2SLS	
		Group 3	Ananat	Group 3	Ananat
Gini index	white	-0.127 (0.090)	**-0.079 (0.037)**	0.020 (-0.118)	**-0.334 (0.099)**
	Black	**-0.448 (0.136)**	**0.459 (0.093)**	-0.127 (0.090)	**-0.875 (0.409)**
Poverty rate	white	**-0.094 (0.024)**	**-0.073 (0.019)**	-0.104 (0.064)	**-0.196 (0.065)**
	Black	**0.269 (0.054)**	**0.182 (0.045)**	**0.383 (0.150)**	**0.258 (0.108)**
90 white: 90 Black		0.193 (0.408)	0.111 (0.086)	-0.250 (0.283)	-0.131 (0.312)
10 white: 10 Black		0.408 (0.305)	**1.295 (0.249)**	1.240 (0.779)	**2.727 (0.867)**
90 white: 10 Black		0.309 (0.350)	**1.172 (0.282)**	1.411 (0.887)	**1.789 (0.756)**
90 Black: 10 white		**-0.292 (0.144)**	-0.234 (0.131)	0.422 (0.374)	**-0.807 (0.384)**

Robustness Checks

A. Table 4A: Results Controlling for 1990 and 1920 City Characteristics

Bolded cells indicate coefficients significantly different from zero at the 0.05 level.

		Outcome: Gini Index				Outcome: Poverty rates				Grp03 N
		Whites		Blacks		Whites		Blacks		
		Ananat	Grp03	Ananat	Grp03	Ananat	Grp03	Ananat	Grp03	
1990	Population	**-0.371** **(0.107)**	**-2.051** **(0.069)**	**0.898** **(0.434)**	**1.393** **(0.055)**	**-0.212** **(0.068)**	**-0.177** **(0.011)**	**0.291** **(0.109)**	**0.538** **(0.012)**	120
	Percent Black	**-0.473** **(0.171)**	**-2.339** **(0.086)**	0.886 (0.547)	**1.491** **(0.079)**	**-0.241** **(0.097)**	**-0.208** **(0.014)**	**0.360** **(0.141)**	**0.583** **(0.028)**	120
	Education	**-0.361** **(0.148)**	0.038 (0.449)	0.887 (0.664)	0.189 (0.633)	**-0.162** **(0.080)**	-0.015 (0.112)	0.222 (0.174)	**0.647** **(0.315)**	120
	Share in manufacturing	**−0.359** **(0.175)**	**-2.585** **(0.286)**	1.106 (0.777)	**1.931** **(0.230)**	**-0.272** **(0.124)**	**-0.313** **(0.044)**	0.219 (0.195)	**0.610** **(0.085)**	110
	Labor force participation	**-0.295** **(0.092)**	-0.412 (0.336)	**0.907** **(0.393)**	0.336 (0.392)	**-0.142** **(0.040)**	-0.073 (0.069)	**0.321** **(0.105)**	**0.651** **(0.172)**	120
	# of local gov'ts (Ananat: N=69)	-0.386 (0.203)	**-1.945** **(0.070)**	**0.792** **(0.277)**	**1.411** **(0.077)**	-0.118 (0.077)	**-0.148** **(0.010)**	**0.519** **(0.169)**	**0.485** **(0.026)**	69
1920	Population	**-0.374** **(0.106)**	**-2.136** **(0.075)**	**0.900** **(0.442)**	**1.415** **(0.063)**	**-0.214** **(0.071)**	**-0.192** **(0.011)**	**0.281** **(0.115)**	**0.552** **(0.023)**	120
	Percent Black	**-0.364** **(0.114)**	**-2.055** **(0.066)**	**0.896** **(0.434)**	**1.449** **(0.055)**	**-0.199** **(0.069)**	**-0.177** **(0.010)**	**0.296** **(0.109)**	**0.535** **(0.021)**	120
	Literacy	**-0.312** **(0.107)**	0.736 (0.562)	**1.028** **(0.469)**	-0.221 (0.888)	**-0.164** **(0.061)**	-0.059 (0.105)	**0.270** **(0.124)**	**0.689** **(0.310)**	80
	Share in manufacturing	**-0.398** **(0.129)**	**-1.351** **(0.226)**	**0.900** **(0.369)**	**0.667** **(0.234)**	**-0.212** **(0.080)**	**-0.169** **(0.039)**	**0.304** **(0.121)**	**0.687** **(0.095)**	49
	Labor force participation	**-0.304** **(0.084)**	-0.372 (0.301)	**0.848** **(0.369)**	0.230 (0.502)	**-0.187** **(0.061)**	-0.108 (0.077)	**0.243** **(0.104)**	**0.647** **(0.202)**	49
	Propensity score	**-0.412** **(0.181)**	**-1.644** **(0.236)**	1.038 (0.639)	**1.036** **(0.241)**	**-0.189** **(0.094)**	**-0.150** **(0.036)**	0.304 (0.177)	**0.517** **(0.091)**	46

Reanalysis

A. Table 5A: 17 MSAs included in Ananat's data but missing from our data

String identifier	City	MSA determined by published Census reports[9]
beaverpa	Beaver, PA	Beaver County, PA PMSA
burlinvt	Burlington, VT	Burlington, VT MSA
elmirany	Elmira, NY	Elmira, NY MSA
fitchbma	Fitchburg, MA	Fitchburg-Leominster, MA MSA
glensfny	Glens Falls, NY	Glens Falls, NY MSA
grandfnd	Grand Forks, ND	Grand Forks, ND MSA
iowaciia	Iowa City, IA	Iowa City, IA MSA
kankakil	Kankakee, IL	Kankakee, IL MSA
lawtonok	Lawton, OK	Lawton, OK MSA
middlect	Middletown, CT	Middletown, CT PMSA
muskegmi	Muskegon, MI	Muskegon, MI MSA
norwalct	Norwalk, CT	Norwalk, CT PMSA
pittsfma	Pittsfield, MA	Pittsfield, MA MSA
portlame	Portland, ME	Portland, ME MSA
portsmnh	Portsmouth, NH	Portsmouth-Dover-Rochester, NH-ME MSA
poughkny	Poughkeepsie, NY	Poughkeepsie, NY MSA
steubeoh	Steubenville, OH	Steubenville-Weirton, OH-WV MSA

[9] We assumed Ananat's level of analysis to be the MSA, which is the most frequently referenced unit of analysis. However, it may be that Ananat splits multi-city MSAs into smaller PMSA (Primary MSA) or CMSA (Consolidated MSA) units. Even so, she does not explicitly state which observations in the data are on PMSA or CMSA levels, so crosswalking occurred in the hand-collection of data (e.g., poverty levels hand-collected with the identifier 'beaverpa' was originally listed in Census publications under 'Beaver County, PA PMSA.' In the IPUMS data, however, no such granularity exists, so no data at the PMSA or CMSA level was collected.)

B. Table 5B: Difference in means using Ananat's original data

The difference in means is between the characteristics and outcomes of MSAs we were able to collect data for (n=104) and those we could not (n=17), using Ananat's original ICPSR data. Bolded cells indicate that sample means differed significantly for those characteristics at a 0.05 level.

Variable		Difference in means
1990 dissimilarity index		-0.0320
RDI		-0.0377
Length of track per square kilometer		**-0.000352**
Gini index	white	-0.0167
	Black	-0.0320
Poverty rate	white	-0.00753
	Black	-0.00681
90 white: 90 Black		-0.0207
10 white: 10 Black		-0.137
90 white: 10 Black		-0.167
90 Black: 10 white		-0.00884

C Table 5C: Reanalysis of main results, accounting for the 17 missing cities

Bolded cells indicate significant results at the 0.05 level. The shaded column ("Ananat") and column titled "Group 3" have been copied over from our main results, seen in Table 3A.

Outcome	Race	OLS				2SLS			
		Ananat	Ananat - 17	Group 3 + 17	Group 3	Ananat	Ananat - 17	Group 3 + 17	Group 3
n		121	104	121	104	121	104	121	104
Gini index	white	**-0.079** **(0.034)**[10]	**-0.091** **(0.037)**	-0.155 (0.091)	-0.127 (0.090)	**-0.334** **(0.113)**	**-0.343** **(0.115)**	-0.103 (0.248)	0.020 (-0.118)
	Black	**0.459** **(0.102)**	**0.494** **(0.097)**	**-0.347** **(0.132)**	**-0.448** **(0.136)**	**0.875** **(0.302)**	**0.877** **(0.272)**	-0.046 (0.365)	-0.127 (0.090)
Poverty rate	white	**-0.073** **(0.022)**	**-0.078** **(0.025)**	**-0.087** **(0.022)**	**-0.094** **(0.024)**	**-0.196** **(0.07)**	**-0.197** **(0.072)**	-0.109 (0.059)	-0.104 (0.064)
	Black	**0.182** **(0.051)**	**0.189** **(0.055)**	**0.256** **(0.051)**	**0.269** **(0.054)**	0.258 (0.144)	0.246 (0.145)	**0.401** **(0.143)**	**0.383** **(0.150)**
90 white: 90 Black		0.111 (0.101)	0.075 (0.106)	**0.212** **(0.104)**	0.193 (0.408)	-0.131 (0.287)	-0.161 (0.282)	-0.212 (0.302)	-0.250 (0.283)
10 white: 10 Black		**1.295** **(0.272)**	**1.393** **(0.291)**	0.432 (0.282)	0.408 (0.305)	**2.727** **(0.836)**	**2.813** **(0.840)**	1.222 (0.785)	1.240 (0.779)
90 white: 10 Black		**1.172** **0.286**	**1.258** **(0.304)**	0.450 (0.348)	0.309 (0.350)	**1.789** **(0.808)**	**1.882** **(0.808)**	1.593 (0.983)	1.411 (0.887)
90 Black: 10 white		-0.234 (0.132)	-0.209 (0.145)	-0.195 (0.185)	**-0.292** **(0.144)**	**-0.807** **(0.394)**	-0.770 (0.404)	0.583 (0.541)	0.422 (0.374)

[10] Heteroskedasticity robust standard errors as formulated in Davidson and MacKinnon, (1993), commonly called "HC3" errors, are reported in parentheses. The standard errors reported in this table for Ananat's original regression are slightly different than what is reported in her original table, which as far as we can tell, used Stata's default "robust" option and are as formulated in MacKinnon and White (1985) (i.e., referred to as "HC1" errors). We do not replicate her standard errors faithfully in this table to facilitate comparison across regressions.

D Table 5D: Lin's estimator with various choices of covariates

Bolded cells indicate significant effects at the 0.05 level.

Covariates	N	Poverty Rates		Gini Index		Income Percentile Ratios			
		white	Black	white	Black	90w:90b	10w:10b	90w:10b	90b:10w
Ananat's 2SLS results	121	**-0.196 (0.07)**	0.258 (0.144)	**-0.334 (0.113)**	**0.875 (0.302)**	-0.131 (0.287)	**2.727 (0.836)**	**1.789 (0.808)**	**-0.807 (0.394)**
Track density	104	**-0.023 (0.008)**	0.039 (0.019)	-0.046 (0.034)	**-0.104 (0.040)**	0.067 (0.044)	0.029 (0.126)	-0.014 (0.135)	**-0.111 (0.049)**
Propensity score[11]	46	0.0007 (0.009)	**0.052 (0.020)**	0.072 (0.039)	-0.075 (0.092)	0.081 (0.056)	-0.044 (0.361)	0.070 (0.294)	0.034 (0.147)
Population in 1990	104	**-0.021 (0.006)**	**0.064 (0.016)**	-0.032 (0.027)	-0.063 (0.035)	0.043 (0.029)	0.104 (0.082)	0.069 (0.093)	**-0.078 (0.039)**
% Black in 1990	104	**-0.015 (0.006)**	**0.061 (0.016)**	**-0.085 (0.029)**	-0.047 (0.036)	0.036 (0.029)	**0.197 (0.091)**	0.124 (0.102)	-0.109 (0.061)
Pop. and % Black in 1990	104	**-0.013 (0.006)**	**0.067 (0.016)**	**-0.078 (0.029)**	-0.029 (0.038)	0.032 (0.029)	**0.209 (0.089)**	0.139 (0.101)	-0.102 (0.061)
Population in 1920	104	**-0.015 (0.006)**	**0.059 (0.016)**	-0.038 (0.038)	-0.064 (0.034)	0.045 (0.029)	0.114 (0.089)	0.113 (0.108)	-0.046 (0.042)
% Black in 1920	84	**-0.018 (0.008)**	**0.053 (0.109)**	-0.047 (0.044)	-0.071 (0.042)	0.042 (0.030)	0.172 (0.101)	0.172 (0.120)	-0.043 (0.050)
Pop. and % Black in 1990, 1920	84	-0.004 (0.008)	**0.070 (0.017)**	-0.049 (0.050)	-0.009 (0.041)	0.035 (0.039)	**0.256 (0.142)**	0.251 (0.142)	-0.039 (0.068)
Education in 1990[12]	104	-0.079 (1.724)	-1.185 (1.504)	11.206 (6.659)	-2.080 (4.487)	0.560 (7.978)	-10.109 (38.476)	-11.335 (34.338)	-1.786 (15.087)
Educ., Pop., and % Black in 1990, 1920	84	1.151 (2.950)	0.131 (4.283)	14.254 (11.41)	0.990 (8.089)	1.053 (15.71)	-4.763 (63.924)	-2.251 (47.374)	1.459 (35.027)

[11] The propensity score model of the probability of having an above-median Railroad Division Index based on 1920 city characteristics (population, percent Black, percent of labor force in manufacturing, literacy rate, labor force participation, and distance from the South)

[12] This includes the percentages of high school dropouts, high school graduates, college dropouts, and college graduates among the white and Black populations of the MSA.

Code

Assessing the Assumptions (R):

```
library(car)
library(tidyverse)

## load data and change units to match Ananat
dat <- read.csv("ourdata.csv", row.names = 1)
dat$povrate_w <- dat$povrate_w / 100
dat$povrate_b <- dat$povrate_b / 100
dat$lnw10b90 <- dat$lnw10b90 * -1
dat$area1910.y <- dat$area1910.y / 1000
dat$passpc1910 <- dat$passpc1910 / 1000
dat$pop1910 <- dat$pop1910 / 1000
dat$incseg.y <- replace(dat$incseg.y, dat$incseg.y == ".", NA) %>%
  as.numeric()

## Assumption 1
asspt1 <- lm(dism1990 ~ herf + lenper, dat)
asspt1$coefficients[2]
sqrt(abs(hccm(asspt1)))

## Assumption 2

# physical area
asspt2a <- lm(area1910.y ~ herf + lenper, dat)
asspt2a$coefficients[2]
sqrt(abs(hccm(asspt2a)))

# population
asspt2b <- lm(pop1910 ~ herf + lenper, dat)
asspt2b$coefficients[2]
sqrt(abs(hccm(asspt2b)))

# ethnic dissm
asspt2c <- lm(dism1910 ~ herf + lenper, dat)
asspt2c$coefficients[2]
sqrt(abs(hccm(asspt2c)))

# ethnic isolation
asspt2d <- lm(isol1910 ~ herf + lenper, dat)
asspt2d$coefficients[2]
sqrt(abs(hccm(asspt2d)))

# percent Black
perB <- dat$b1910 / dat$pop1910
asspt2e <- lm(perB ~ herf + lenper, dat)
asspt2e$coefficients[2]
```

```
sqrt(abs(hccm(asspt2e)))

# streetcars per capita
asspt2f <- lm(passpc1910 ~ herf + lenper, dat)
asspt2f$coefficients[2]
sqrt(abs(hccm(asspt2f)))

# income segregation (1990)
asspt2e <- lm(incseg.y ~ herf + lenper, dat)
asspt2e$coefficients[2]
sqrt(abs(hccm(asspt2e)))
```

*Generating propensity scores*data <- read_csv("robustness_checks_data.csv")

```
# find which MSAs had RDIs (column name: ‘herf’) above the median
above_med_herf <- data["herf"] >= median(data[["herf"]])
data["above_med"] <- above_med_herf

# fit the propensity score model
ps_mod <- glm(above_med_herf ~ black1920 + count1920 +
                              ctymanuf_wkrs1920 + ctyliterate1920 +
                              lfp1920 + closeness,
                          data = data, family = "binomial")

# add the propensity scores back into the data
ps <- data.frame(ps = ps_mod$fitted.values,
                 X1 = names(ps_mod$fitted.values)) %>%
         type_convert(ps, col_types = "dd")
data <- data %>% left_join(ps)
```

Finding Lin’s estimator for various covariates

```
# create dichotomized treatment indicator
data <- data %>% slice(1:121) %>% mutate(
    dich_seg = as.numeric(dism1990 >= median(dism1990))
)

outcomes <- c("povrate_w.y", "povrate_b.y",
              "lnwgini.y", "lnbgini.y",
              "lnw90b90", "lnw10b10", "lnw90b10", "lnb90w10")

# centering covariates
cols <- c("lenper.x", "lenper.y", "ps", "pop1990_msa",
                    "pctbk1990_msa", "count1920", "black1920",
                    "hsgrad_w", "hsgrad_b", "hsdrop_w",
                    "hsdrop_b", "somecoll_b", "somecoll_w",
                    "collgrad_w", "collgrad_b", "ngov62")
data[cols] <- apply(data[cols], 2,
                    function(x) x - mean(x, na.rm = T))
```

```
## adjusting for lenper: all other adjustments follow this pattern, replacing
'lenper.y' in the for loop model with the correct covariates

res_lenper <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
      mod <- lm(data[[i]] ~ dich_seg * (lenper.y), data)
      est <- coefficients(mod)[2]
      se <- sqrt(hccm(mod)[2, 2])
      n <- nobs(mod)
      res_lenper <- rbind(res_lenper, c(est, se, n))
}

rownames(res_lenper) = outcomes
colnames(res_lenper) = c("est", "se", "n")
res_lenper$sign <- !(((res_lenper$est - 1.96*res_lenper$se) <= 0) &
                     ((res_lenper$est + 1.96*res_lenper$se) >= 0))

res_lenper

## adjusting for propensity score
res_ps <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (ps), data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_ps <- rbind(res_ps, c(est, se, n))
}

rownames(res_ps) = outcomes
colnames(res_ps) = c("est", "se", "n")
res_ps$sign <- !(((res_ps$est - 1.96*res_ps$se) <= 0) & ((res_ps$est +
1.96*res_ps$se) >= 0))

res_ps

## controlling for 1990 population & 1990 percent Black
res_pop90 <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (pop1990_msa + pctbk1990_msa), data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_pop90 <- rbind(res_pop90, c(est, se, n))
}
```

```
rownames(res_pop90) = outcomes
colnames(res_pop90) = c("est", "se", "n")
res_pop90$sign <- !(((res_pop90$est - 1.96*res_pop90$se) <= 0) &
((res_pop90$est + 1.96*res_pop90$se) >= 0))

res_pop90

## adjusting for population in 1990 and 1920 and percent Black in 1990 and 1920
res_pop <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (pop1990_msa + pctbk1990_msa + count1920 +
black1920),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_pop <- rbind(res_pop, c(est, se, n))
}

rownames(res_pop) = outcomes
colnames(res_pop) = c("est", "se", "n")
res_pop$sign <- !(((res_pop$est - 1.96*res_pop$se) <= 0) & ((res_pop$est +
1.96*res_pop$se) >= 0))

res_pop

## adjusting for education variables
res_ed <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (hsdrop_w + hsdrop_b + hsgrad_w + hsgrad_b +
somecoll_w + somecoll_b + collgrad_w + collgrad_b),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_ed <- rbind(res_ed, c(est, se, n))
}

rownames(res_ed) = outcomes
colnames(res_ed) = c("est", "se", "n")
res_ed$sign <- !(((res_ed$est - 1.96*res_ed$se) <= 0) & ((res_ed$est +
1.96*res_ed$se) >= 0))

res_ed

## adjusting for education and population variables
```

```
res_big <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (hsdrop_w + hsdrop_b + hsgrad_w + hsgrad_b +
somecoll_w + somecoll_b + collgrad_w + collgrad_b + pop1990_msa + pctbk1990_msa
+ count1920 + black1920),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_big <- rbind(res_big, c(est, se, n))
}

rownames(res_big) = outcomes
colnames(res_big) = c("est", "se", "n")
res_big$sign <- !(((res_big$est - 1.96*res_big$se) <= 0) & ((res_big$est +
1.96*res_big$se) >= 0))

res_big

## adjusting only for 1990 population
res_p90 <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (pop1990_msa),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_p90 <- rbind(res_p90, c(est, se, n))
}

rownames(res_p90) = outcomes
colnames(res_p90) = c("est", "se", "n")
res_p90$sign <- !(((res_p90$est - 1.96*res_p90$se) <= 0) & ((res_p90$est +
1.96*res_p90$se) >= 0))

res_p90

## adjusting only for percent Black in 1990
res_pb90 <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (pctbk1990_msa),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_pb90 <- rbind(res_pb90, c(est, se, n))
}

rownames(res_pb90) = outcomes
colnames(res_pb90) = c("est", "se", "n")
res_pb90$sign <- !(((res_pb90$est - 1.96*res_pb90$se) <= 0) & ((res_pb90$est +
1.96*res_pb90$se) >= 0))

res_pb90

## adjusting only for 1920 population
res_p20 <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (count1920),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_p20 <- rbind(res_p20, c(est, se, n))
}

rownames(res_p20) = outcomes
colnames(res_p20) = c("est", "se", "n")
res_p20$sign <- !(((res_p20$est - 1.96*res_p20$se) <= 0) & ((res_p20$est +
1.96*res_p20$se) >= 0))

res_p20

## adjusting only for percent Black in 1920
res_pb20 <- data.frame(est = c(), se = c(), n = c())

for (i in outcomes) {
  mod <- lm(data[[i]] ~ dich_seg * (black1920),
            data)
  est <- coefficients(mod)[2]
  se <- sqrt(hccm(mod)[2, 2])
  n <- nobs(mod)
  res_pb20 <- rbind(res_pb20, c(est, se, n))
}

rownames(res_pb20) = outcomes
colnames(res_pb20) = c("est", "se", "n")
res_pb20$sign <- !(((res_pb20$est - 1.96*res_pb20$se) <= 0) & ((res_pb20$est +
1.96*res_pb20$se) >= 0))

res_pb20
```

In [1]:
```
import numpy as np
import pandas as pd
import os

from collections import Counter
import datetime as dt

import statsmodels.api as sm
```

Importing data

In [2]:
```
aej = pd.read_csv("./data/replication/casey_maindata_v3.csv")
```

In [3]:
```
aej.wgini = aej.wgini*100 # get gini indices instead of gini coefficients
aej.bgini = aej.bgini*100

aej.lnwgini = np.log(aej.wgini)
aej.lnbgini = np.log(aej.bgini)
```

In [4]:
```
aej.head()
```

Out[4]:

	msa	num	numw	numb	perw	perb	w10perc	w90perc	b10perc	b90perc	...
0	akronoh	20032	18364	1668	0.916733	0.083267	2553.0	40997.0	1596.0	30000.0	...
1	albanyny	28993	27989	1004	0.965371	0.034629	3000.0	43736.0	1200.0	32288.0	...
2	altoonpa	4667	4645	22	0.995286	0.004714	2500.0	32000.0	2400.0	33000.0	...
3	anaheica	69648	68385	1263	0.981866	0.018134	4000.0	60500.0	3700.0	44897.0	...
4	annarbmi	8024	7320	704	0.912263	0.087737	2400.0	50000.0	2000.0	38561.0	...

5 rows × 27 columns

In [5]:
```
aej.columns
```

Out[5]:
```
Index(['msa', 'num', 'numw', 'numb', 'perw', 'perb', 'w10perc', 'w90perc',
       'b10perc', 'b90perc', 'w90b90', 'lnw90b90', 'w10b10', 'lnw10b10',
       'w90b10', 'lnw90b10', 'b90w10', 'lnb90w10', 'wgini', 'bgini', 'lnwgin
i',
       'lnbgini', 'dism1990', 'herf', 'lenper', 'povrate_w', 'povrate_b'],
      dtype='object')
```

First stage

$$\text{Segregation} = \beta_0 + \beta_1 \cdot \text{RDI} + \beta_2 \cdot \text{Railroad length}$$

In [6]:
```
aej.columns
```

Out[6]:
```
Index(['msa', 'num', 'numw', 'numb', 'perw', 'perb', 'w10perc', 'w90perc',
       'b10perc', 'b90perc', 'w90b90', 'lnw90b90', 'w10b10', 'lnw10b10',
       'w90b10', 'lnw90b10', 'b90w10', 'lnb90w10', 'wgini', 'bgini', 'lnwgin
i',
       'lnbgini', 'dism1990', 'herf', 'lenper', 'povrate_w', 'povrate_b'],
      dtype='object')
```

In [7]:
```
aej.shape
```

Out[7]:
```
(104, 27)
```

```
In [8]: Y = aej[["dism1990"]]
        X = aej[["herf","lenper"]]
        X = sm.add_constant(X)

        mod = sm.OLS(Y, X, cov_type='HC3', hasconst = 1)
        res = mod.fit()
        print("Regressing segregation measure (dissimilarity index) on railroad divisi
        on index (RDI) and length of railroad track")
        print(res.summary())
```

```
Regressing segregation measure (dissimilarity index) on railroad division ind
ex (RDI) and length of railroad track
                            OLS Regression Results                            
=============================================================================
=
Dep. Variable:               dism1990   R-squared:                       0.23
0
Model:                            OLS   Adj. R-squared:                  0.21
5
Method:                 Least Squares   F-statistic:                     15.0
7
Date:                Fri, 06 Nov 2020   Prob (F-statistic):           1.87e-0
6
Time:                        16:49:00   Log-Likelihood:                 74.10
8
No. Observations:                 104   AIC:                            -142.
2
Df Residuals:                     101   BIC:                            -134.
3
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
const          0.2816      0.061      4.646      0.000       0.161       0.40
2
herf           0.3755      0.084      4.490      0.000       0.210       0.54
1
lenper        17.9535      9.061      1.981      0.050      -0.021      35.92
8
=============================================================================
=
Omnibus:                       11.522   Durbin-Watson:                   1.56
3
Prob(Omnibus):                  0.003   Jarque-Bera (JB):                3.90
8
Skew:                          -0.061   Prob(JB):                        0.14
2
Kurtosis:                       2.058   Cond. No.                         95 
2.
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [9]:
```
aej.dism1990[:10]
```

Out[9]:
```
0    0.692728
1    0.619620
2    0.521674
3    0.345086
4    0.499234
5    0.632075
6    0.511915
7    0.634861
8    0.741388
9    0.516248
Name: dism1990, dtype: float64
```

In [10]:
```
res.fittedvalues[:10]
```

Out[10]:
```
0    0.615171
1    0.540778
2    0.568859
3    0.540377
4    0.590299
5    0.768224
6    0.640780
7    0.561177
8    0.608528
9    0.567242
dtype: float64
```

In [11]:
```
# adding back fitted results
aej["fit_dism1990"] = 0
aej["fit_dism1990"] = res.fittedvalues
```

Creating our version of Ananat's Table 2 (regression results)

In [12]:
```
table = pd.DataFrame(columns = ["dependent_variable",
                               "ols","ols_se", "ols_sig",
                               "tsls", "tsls_se", "tsls_sig",])
```

In [13]:
```
depvars = ['lnwgini', 'lnbgini', 'povrate_w', 'povrate_b',
           'lnw90b90', 'lnw10b10', 'lnw90b10', 'lnb90w10']
```

In [14]:

```
for depvar in depvars:
    # OLS
    Y = aej[depvar]
    X = aej["dism1990"]
    X = sm.add_constant(X)
    mod = sm.OLS(Y, X, cov_type='HC3')
    res = mod.fit()
    olsest, olsse = res.params[1], res.bse[1]
    olssig = 0
    if not (0 > olsest-1.96*olsse and 0 < olsest+1.96*olsse):
        olssig = 1

    # TSLS
    Y = aej[depvar]
    X = aej[["fit_dism1990", "lenper"]]
    X = sm.add_constant(X)
    mod = sm.OLS(Y, X, cov_type='HC3' )
    res = mod.fit()
    tslsest, tslsse = res.params[1], res.bse[1]
    tslssig = 0
    if not (0 > tslsest-1.96*tslsse and 0 < tslsest+1.96*tslsse):
        tslssig = 1


    table = table.append({'dependent_variable':depvar,
                 'ols':olsest,
                 'ols_se': olsse,
                 'ols_sig': olssig,
                 'tsls': tslsest,
                 'tsls_se':tslsse,
                 'tsls_sig': tslssig}, ignore_index=True)
```

In [15]:

```
np.std(aej.dism1990)
```

Out[15]:

```
0.13520943746016845
```

In [16]:

```
table # new table
```

Out[16]:

	dependent_variable	ols	ols_se	ols_sig	tsls	tsls_se	tsls_sig
0	lnwgini	-0.127505	0.090219	0	0.020029	0.233035	0
1	lnbgini	-0.447516	0.135756	1	-0.117879	0.361537	0
2	povrate_w	-0.094240	0.023890	1	-0.103501	0.064271	0
3	povrate_b	0.268971	0.054384	1	0.382992	0.150205	1
4	lnw90b90	0.193009	0.109200	0	-0.250143	0.282791	0
5	lnw10b10	0.408133	0.305404	0	1.239700	0.779026	0
6	lnw90b10	0.309125	0.350111	0	1.411229	0.887306	0
7	lnb90w10	-0.292017	0.144335	1	0.421672	0.374535	0

In [17]:

```
table.to_csv("./data/casey_table2.csv", index = False)
```

Below is the in-depth exploration of each and every regression, but are all the relevant results are generated in table above.

Table 2 Panel 1

OLS column

In [18]:

```
mod = sm.OLS(aej.lnwgini, aej.dism1990, cov_type='HC3')
res = mod.fit()
print("Regressing log white Gini index on segregation index")
print(res.summary())
```

```
Regressing log white Gini index on segregation index
                                 OLS Regression Results                        
=============================================================================
==========
Dep. Variable:                lnwgini   R-squared (uncentered):
0.944
Model:                            OLS   Adj. R-squared (uncentered):
0.943
Method:                 Least Squares   F-statistic:
1727.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):
3.60e-66
Time:                        16:49:01   Log-Likelihood:
-127.60
No. Observations:                 104   AIC:
257.2
Df Residuals:                     103   BIC:
259.8
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
dism1990       5.7460      0.138     41.553      0.000       5.472       6.02
0
=============================================================================
=
Omnibus:                       20.693   Durbin-Watson:                   1.72
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                5.16
5
Skew:                           0.087   Prob(JB):                       0.075
6
Kurtosis:                       1.922   Cond. No.                         1.0
0
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [19]:

```
mod = sm.OLS(aej.lnbgini, aej.dism1990, cov_type='HC3')
res = mod.fit()
print("Regressing log Black Gini index on segregation index")
print(res.summary())
```

```
Regressing log Black Gini index on segregation index
                                 OLS Regression Results                        
=============================================================================
==========
Dep. Variable:                lnbgini   R-squared (uncentered):
0.937
Model:                            OLS   Adj. R-squared (uncentered):
0.937
Method:                 Least Squares   F-statistic:
1543.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):
8.28e-64
Time:                        16:49:01   Log-Likelihood:
-141.47
No. Observations:                 104   AIC:
284.9
Df Residuals:                     103   BIC:
287.6
Df Model:                           1
Covariance Type:            nonrobust
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
dism1990       6.2076      0.158     39.286      0.000       5.894       6.52
1
=============================================================================
=
Omnibus:                        4.861   Durbin-Watson:                   1.67
9
Prob(Omnibus):                  0.088   Jarque-Bera (JB):                2.68
7
Skew:                          -0.147   Prob(JB):                        0.26
1
Kurtosis:                       2.270   Cond. No.                         1.0
0
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [20]:

```
mod = sm.OLS(aej.povrate_w, aej.dism1990, cov_type='HC3')
res = mod.fit()
print("Regressing white poverty rate on segregation index")
print(res.summary())
```

```
Regressing white poverty rate on segregation index
                                 OLS Regression Results                                
=============================================================================
==========
Dep. Variable:              povrate_w   R-squared (uncentered):                   
0.778
Model:                            OLS   Adj. R-squared (uncentered):              
0.776
Method:                 Least Squares   F-statistic:                              
360.6
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                       
2.00e-35
Time:                        16:49:01   Log-Likelihood:                           
170.34
No. Observations:                 104   AIC:                                      
-338.7
Df Residuals:                     103   BIC:                                      
-336.0
Df Model:                           1                                              
Covariance Type:            nonrobust                                              
=============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
-----------------------------------------------------------------------------
-
dism1990       0.1497      0.008     18.991      0.000       0.134       0.16
5
=============================================================================
=
Omnibus:                       25.976   Durbin-Watson:                   2.05
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.92
1
Skew:                           1.101   Prob(JB):                     7.89e-1
0
Kurtosis:                       5.196   Cond. No.                         1.0
0
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [21]:
```python
mod = sm.OLS(aej.povrate_b, aej.dism1990, cov_type='HC3')
res = mod.fit()
print("Regressing black poverty rate on segregation index")
print(res.summary())
```

```
Regressing black poverty rate on segregation index
                                 OLS Regression Results                                
=============================================================================================
==========
Dep. Variable:              povrate_b   R-squared (uncentered):                   
0.929
Model:                            OLS   Adj. R-squared (uncentered):              
0.929
Method:                 Least Squares   F-statistic:                              
1354.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                       
4.45e-61
Time:                        16:49:01   Log-Likelihood:                           
114.06
No. Observations:                 104   AIC:                                      
-226.1
Df Residuals:                     103   BIC:                                      
-223.5
Df Model:                           1                                                  
Covariance Type:            nonrobust                                                  
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
dism1990       0.4983      0.014     36.801      0.000       0.471       0.52
5
==============================================================================
=
Omnibus:                        1.168   Durbin-Watson:                   2.24
0
Prob(Omnibus):                  0.558   Jarque-Bera (JB):                1.15
3
Skew:                          -0.133   Prob(JB):                        0.56
2
Kurtosis:                       2.559   Cond. No.                         1.0
0
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

2SLS column

```
In [22]: mod = sm.OLS(aej[["lnwgini"]], aej[["dism1990"]], cov_type='HC3' )
         res = mod.fit()
         #print("Regressing segregation measure (dissimilarity index) on railroad divis
         ion index (RDI) and length of railroad track")
         print(res.summary())
```

```
                                 OLS Regression Results
==============================================================================
==========
Dep. Variable:                lnwgini   R-squared (uncentered):
0.944
Model:                            OLS   Adj. R-squared (uncentered):
0.943
Method:                 Least Squares   F-statistic:
1727.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):
3.60e-66
Time:                        16:49:01   Log-Likelihood:
-127.60
No. Observations:                 104   AIC:
257.2
Df Residuals:                     103   BIC:
259.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
dism1990       5.7460      0.138     41.553      0.000       5.472       6.02
0
==============================================================================
=
Omnibus:                       20.693   Durbin-Watson:                   1.72
5
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                5.16
5
Skew:                           0.087   Prob(JB):                       0.075
6
Kurtosis:                       1.922   Cond. No.                        1.0
0
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [23]:

```
mod = sm.OLS(aej[["lnwgini"]], aej[["fit_dism1990", "lenper"]], cov_type='HC3'
)
res = mod.fit()
#print("Regressing segregation measure (dissimilarity index) on railroad divis
ion index (RDI) and length of railroad track")
print(res.summary())
```

```
                                 OLS Regression Results                        
==============================================================================
==========
Dep. Variable:                lnwgini   R-squared (uncentered):               
0.990
Model:                            OLS   Adj. R-squared (uncentered):          
0.989
Method:                 Least Squares   F-statistic:                          
4818.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                   
1.07e-101
Time:                        16:49:01   Log-Likelihood:                       
-40.155
No. Observations:                 104   AIC:                                  
84.31
Df Residuals:                     102   BIC:                                  
89.60
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
fit_dism1990     6.2694      0.079     79.792      0.000       6.114       6.
425
lenper        -151.4882     27.503     -5.508      0.000    -206.039     -96.
937
==============================================================================
=
Omnibus:                       13.683   Durbin-Watson:                   1.87
6
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               15.15
9
Skew:                           0.923   Prob(JB):                     0.00051
1
Kurtosis:                       3.306   Cond. No.                         44
9.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [24]:

```
mod = sm.OLS(aej[["lnbgini"]], aej[["fit_dism1990", "lenper"]], cov_type='HC3'
)
res = mod.fit()
#print("Regressing segregation measure (dissimilarity index) on railroad divis
ion index (RDI) and length of railroad track")
print(res.summary())
```

```
                                 OLS Regression Results                         
==============================================================================
==========
Dep. Variable:                lnbgini   R-squared (uncentered):               
0.988
Model:                            OLS   Adj. R-squared (uncentered):          
0.987
Method:                 Least Squares   F-statistic:                          
4092.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                   
4.03e-98
Time:                        16:49:01   Log-Likelihood:                       
-56.935
No. Observations:                 104   AIC:                                  
117.9
Df Residuals:                     102   BIC:                                  
123.2
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
fit_dism1990     6.8376      0.092     74.057      0.000       6.654       7.
021
lenper        -193.2125     32.318     -5.978      0.000    -257.315    -129.
110
==============================================================================
=
Omnibus:                        2.201   Durbin-Watson:                   2.03
2
Prob(Omnibus):                  0.333   Jarque-Bera (JB):                1.80
0
Skew:                           0.317   Prob(JB):                        0.40
7
Kurtosis:                       3.110   Cond. No.                         44
9.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [25]:

```
mod = sm.OLS(aej[["povrate_w"]], aej[["fit_dism1990", "lenper"]], cov_type='HC
3' )
res = mod.fit()
#print("Regressing segregation measure (dissimilarity index) on railroad divis
ion index (RDI) and length of railroad track")
print(res.summary())
```

```
                                 OLS Regression Results                        
==============================================================================
==========
Dep. Variable:              povrate_w   R-squared (uncentered):               
0.860
Model:                            OLS   Adj. R-squared (uncentered):          
0.858
Method:                 Least Squares   F-statistic:                          
314.0
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                   
2.55e-44
Time:                        16:49:02   Log-Likelihood:                       
194.46
No. Observations:                 104   AIC:                                  
-384.9
Df Residuals:                     102   BIC:                                  
-379.6
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
------------------------------------------------------------------------------
---
fit_dism1990     0.1736      0.008     21.088      0.000       0.157       0.
190
lenper          -7.7109      2.882     -2.676      0.009     -13.427      -1.
995
==============================================================================
=
Omnibus:                       40.630   Durbin-Watson:                   2.32
8
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               106.68
9
Skew:                           1.423   Prob(JB):                     6.80e-2
4
Kurtosis:                       7.064   Cond. No.                        44
9.
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

In [26]:

```
mod = sm.OLS(aej[["povrate_b"]], aej[["fit_dism1990", "lenper"]], cov_type='HC
3' )
res = mod.fit()
#print("Regressing segregation measure (dissimilarity index) on railroad divis
ion index (RDI) and length of railroad track")
print(res.summary())
```

```
                                 OLS Regression Results                          
=============================================================================
==========
Dep. Variable:              povrate_b   R-squared (uncentered):             
0.930
Model:                            OLS   Adj. R-squared (uncentered):        
0.929
Method:                 Least Squares   F-statistic:                        
677.9
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                 
1.23e-59
Time:                        16:49:02   Log-Likelihood:                     
114.59
No. Observations:                 104   AIC:                                
-225.2
Df Residuals:                     102   BIC:                                
-219.9
Df Model:                           2                                       
Covariance Type:            nonrobust                                       
=============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
-----------------------------------------------------------------------------
---
fit_dism1990     0.5279      0.018     29.746      0.000       0.493       0.
563
lenper         -10.8623      6.211     -1.749      0.083     -23.182       1.
458
=============================================================================
=
Omnibus:                        1.074   Durbin-Watson:                   1.83
1
Prob(Omnibus):                  0.584   Jarque-Bera (JB):                1.15
3
Skew:                           0.179   Prob(JB):                        0.56
2
Kurtosis:                       2.628   Cond. No.                         44
9.
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

Table 2 Panel 2

In [27]:
```
depvars
```

Out[27]:
```
['lnwgini',
 'lnbgini',
 'povrate_w',
 'povrate_b',
 'lnw90b90',
 'lnw10b10',
 'lnw90b10',
 'lnb90w10']
```

In [28]:

```
# TSLS
print(depvars[2])
Y = aej[depvars[2]]
X = aej[["fit_dism1990", "lenper"]]
X = sm.add_constant(X)
mod = sm.OLS(Y, X, cov_type='HC3' )
res = mod.fit()
print(res.summary())
```

```
povrate_w
                            OLS Regression Results                            
=============================================================================
=
Dep. Variable:              povrate_w   R-squared:                       0.04
4
Model:                            OLS   Adj. R-squared:                  0.02
5
Method:                 Least Squares   F-statistic:                     2.32
6
Date:                Fri, 06 Nov 2020   Prob (F-statistic):              0.10
3
Time:                        16:49:02   Log-Likelihood:                 203.3
6
No. Observations:                 104   AIC:                            -400.
7
Df Residuals:                     101   BIC:                            -392.
8
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
=============================================================================
===
                   coef    std err          t      P>|t|      [0.025      0.9
75]
-----------------------------------------------------------------------------
---
const            0.1534      0.035      4.342      0.000       0.083       0.
223
fit_dism1990    -0.1035      0.064     -1.610      0.110      -0.231       0.
024
lenper          -0.7650      3.103     -0.247      0.806      -6.920       5.
390
=============================================================================
=
Omnibus:                       33.089   Durbin-Watson:                   2.19
0
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               76.70
0
Skew:                           1.190   Prob(JB):                     2.21e-1
7
Kurtosis:                       6.469   Cond. No.                     1.05e+0
3
=============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.05e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

In [ ]:

Filling in Table 2 Panel 2

In [29]:
```
mod = sm.OLS(aej.lnb90w10, aej.dism1990, cov_type='HC3')
res = mod.fit()
print("Regressing black poverty rate on segregation index")
print(res.summary())
```

```
Regressing black poverty rate on segregation index
                                 OLS Regression Results                      
==============================================================================
==========
Dep. Variable:               lnb90w10   R-squared (uncentered):              
0.933
Model:                            OLS   Adj. R-squared (uncentered):         
0.933
Method:                 Least Squares   F-statistic:                         
1443.
Date:                Fri, 06 Nov 2020   Prob (F-statistic):                  
2.14e-62
Time:                        16:49:02   Log-Likelihood:                      
-98.773
No. Observations:                 104   AIC:                                 
199.5
Df Residuals:                     103   BIC:                                 
202.2
Df Model:                           1                                        
Covariance Type:            nonrobust                                        
==============================================================================
=
                 coef    std err          t      P>|t|      [0.025      0.97
5]
------------------------------------------------------------------------------
-
dism1990       3.9809      0.105     37.982      0.000       3.773       4.18
9
==============================================================================
=
Omnibus:                       14.083   Durbin-Watson:                   1.67
8
Prob(Omnibus):                  0.001   Jarque-Bera (JB):                5.80
2
Skew:                           0.327   Prob(JB):                       0.055
0
Kurtosis:                       2.046   Cond. No.                        1.0
0
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
```

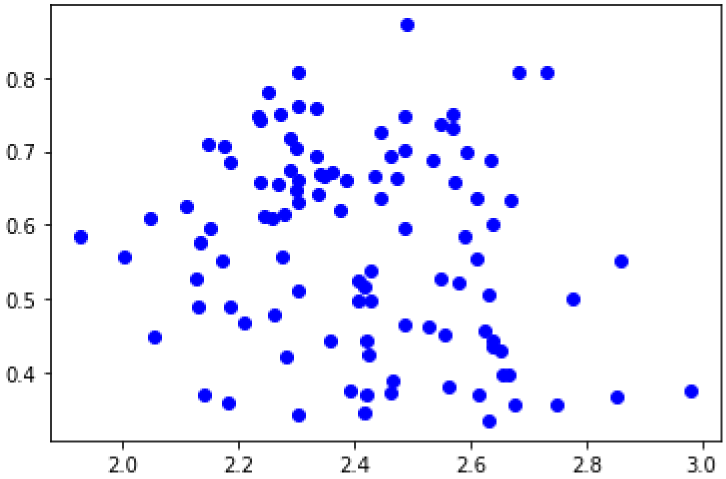
In [30]:
```
import matplotlib.pyplot as plt
```

In [31]:
```
%matplotlib inline

df = aej
x = df.lnb90w10
y = df.dism1990

#plt.xlim(0, 5)
plt.plot(x,y, "bo")
```

Out[31]:
```
[<matplotlib.lines.Line2D at 0x29529f6a748>]
```



In [32]:
```
aej[["lnb90w10", "dism1990"]]
```

Out[32]:

	lnb90w10	dism1990
0	2.463928	0.692728
1	2.376083	0.619620
2	2.580217	0.521674
3	2.418077	0.345086
4	2.776773	0.499234
...	...	...
99	2.407946	0.524915
100	2.556853	0.451906
101	2.148434	0.709774
102	2.484907	0.748709
103	2.657341	0.397027

104 rows × 2 columns

In [ ]:

In [1]:
```
import pandas as pd
import os
import numpy as np
```

Merging data to get Table 3 (Ananat's robustness checks)

In [3]:
```
robust = pd.read_csv("./data/ipums/robustness_checks_data_ravenna.csv")
robust = robust.drop(["Unnamed: 0", "lngini_w", "lngini_b"], axis = 1)
```

In [4]:
```
city = pd.read_csv("./data/ipums/1920_city_characteristics.csv")
```

In [5]:
```
rep = pd.read_csv("./data/replication/casey_maindata_v3.csv")
```

In [6]:
```
rep.columns
```

Out[6]:
```
Index(['msa', 'num', 'numw', 'numb', 'perw', 'perb', 'w10perc', 'w90perc',
       'b10perc', 'b90perc', 'w90b90', 'lnw90b90', 'w10b10', 'lnw10b10',
       'w90b10', 'lnw90b10', 'b90w10', 'lnb90w10', 'wgini', 'bgini', 'lnwgi
ni',
       'lnbgini', 'dism1990', 'herf', 'lenper', 'povrate_w', 'povrate_b'],
      dtype='object')
```

In [7]:
```
len(city.METAREAD.unique())
```

Out[7]:
```
111
```

In [8]:
```
len(robust.msafips.unique())
```

Out[8]:
```
121
```

In [10]:
```
# robust.merge(
#     rep[['msa', 'lnwgini','lnbgini']], how = "outer",
#     left_on = "msafips", right_on = "msa")
```

In [11]:
```
crep = robust.merge(city, left_on = "msafips", right_on = "METAREAD", how = "
    rep[['msa', 'lnwgini','lnbgini']], left_on = "name", right_on = "msa", ho
crep.columns
```

Out[11]:
```
Index(['name', 'msafips', 'state', 'county', 'herf', 'lenper', 'closeness',
       'dism1990', 'povrate_w', 'povrate_b', 'pop1990_msa', 'pctbk1990_ms
a',
       'hsdrop_w', 'hsgrad_w', 'somecoll_w', 'collgrad_w', 'hsdrop_b',
       'hsgrad_b', 'somecoll_b', 'collgrad_b', 'manshr', 'lfp_w', 'lfp_b',
       'ngov62', 'count1920', 'black1920', 'percent illiterate',
       'ctyliterate1920', 'lfp1920', 'ctymanuf_wkrs1920', 'ps', 'herfscor
e',
       'METAREAD', 'literate', 'lfp', 'manuf', 'msa', 'lnwgini', 'lnbgin
i'],
      dtype='object')
```

In [12]:
```
crep.shape
```

Out[12]:
```
(183, 39)
```

In [14]:
```
crep[['name', 'state', 'county', 'msafips', 'herf', 'lenper', 'closeness',
       'dism1990', 'lnwgini','lnbgini', 'povrate_w', 'povrate_b',
       'pop1990_msa', 'pctbk1990_msa', 'hsdrop_w', 'hsgrad_w', 'somecoll_w',
       'collgrad_w', 'hsdrop_b', 'hsgrad_b', 'somecoll_b', 'collgrad_b',
       'manshr', 'lfp_w', 'lfp_b', 'ngov62', 'count1920', 'black1920',
       'percent illiterate', 'literate', 'lfp', 'manuf', 'ps',
       'herfscore']].to_csv("./data/ipums/robustness_check_data_ravenna_reg.c
```

In [1]:

```python
import pandas as pd
import os
import numpy as np
```

In [2]:

```python
from statsmodels.sandbox.regression.gmm import IV2SLS
```

Running Table 3 (Ananat's robustness checks) regressions

In [3]:

```python
dat = pd.read_csv("./data/ipums/robustness_check_data_ravenna_reg.csv")
dat = dat.drop("literate", axis = 1)
dat["literate"] = 100 - dat["percent illiterate"]
# dat = dat.drop("percent literate", axis = 1)
dat.head()
```

Out[3]:

	name	state	county	msafips	herf	lenper	closeness	dism1990	lnwgini
0	akronoh	OH	Summit	80.0	0.832052	0.001176	-107.519997	0.692728	-1.295651
1	albanyny	NY	Albany	160.0	0.667636	0.000471	-231.380005	0.619620	-1.299983
2	altoonpa	PA	Blair	280.0	0.726499	0.000804	-80.599998	0.521674	-1.271674
3	anaheica	CA	Orange	360.0	0.671304	0.000372	-680.750000	0.345086	-1.058413
4	annarbmi	MI	Washtenaw	440.0	0.798910	0.000484	-247.300003	0.499234	-1.017579

5 rows × 34 columns

In [4]:

```python
dat["povrate_w"] = dat["povrate_w"]/100
dat["povrate_b"] = dat["povrate_b"]/100
```

In [5]:

```python
dat.columns
```

Out[5]:

```
Index(['name', 'state', 'county', 'msafips', 'herf', 'lenper', 'closeness',
       'dism1990', 'lnwgini', 'lnbgini', 'povrate_w', 'povrate_b',
       'pop1990_msa', 'pctbk1990_msa', 'hsdrop_w', 'hsgrad_w', 'somecoll_
w',
       'collgrad_w', 'hsdrop_b', 'hsgrad_b', 'somecoll_b', 'collgrad_b',
       'manshr', 'lfp_w', 'lfp_b', 'ngov62', 'count1920', 'black1920',
       'percent illiterate', 'lfp', 'manuf', 'ps', 'herfscore', 'literat
e'],
      dtype='object')
```

In [6]:

```python
dat.shape
```

Out[6]:

```
(183, 34)
```

In [7]:
```
dat.dropna().shape
```

Out[7]:
```
(36, 34)
```

In [8]:
```
ys = ['lnwgini', 'lnbgini', 'povrate_w', 'povrate_b']
```

In [9]:
```
xs = ['pop1990_msa', 'pctbk1990_msa',"education",'manshr', 'lfp1990', "ngov62
      'count1920', 'black1920','literate', 'manuf', 'lfp', "ps"]
```

In [10]:
```
reg = []
```

In [11]:

```
for x in xs:
    coefs = []
    ses = []
    ps = []

    for y in ys:
#         print(y)
#         print(x)
        if x == "education":
            edu_cols = ['hsdrop_w', 'hsgrad_w', 'somecoll_w', 'collgrad_w', 'hsdrop_b', 'hsgrad_b', 'somecoll_b', 'collgrad_b',]
            cols = [y, "dism1990", "lenper", "herf"] + edu_cols

            df = dat[cols].dropna()

            mod = IV2SLS(
                endog = df[y],
                exog = df[["dism1990", "lenper"] + edu_cols],
                instrument = df[["herf", "lenper"] + edu_cols])
            res = mod.fit()

        elif x == "lfp1990":
            lfp_cols = ['lfp_w', 'lfp_b',]
            cols = [y, "dism1990", "lenper", "herf"] + lfp_cols
            df = dat[cols].dropna()

            mod = IV2SLS(
                endog = df[y],
                exog = df[["dism1990", "lenper"] + lfp_cols],
                instrument = df[["herf", "lenper"] + lfp_cols])
            res = mod.fit()

        else:
            cols = [y, x, "dism1990", "lenper", "herf"]
            df = dat[cols].dropna()

            mod = IV2SLS(
                endog = df[y],
                exog = df[["dism1990", x, "lenper"]],
                instrument = df[["herf", x, "lenper"]])
            res = mod.fit()

        coefs.append(res.params[0])
        ses.append(res.bse[0])
        ps.append(res.pvalues[0])
#         print("\n")

    reg.append({
        "ctrl": x,
        "type": "est",
        ys[0]: coefs[0],
        ys[1]: coefs[1],
        ys[2]: coefs[2],
        ys[3]: coefs[3],
    })
    reg.append({
        "ctrl": x,
        "type": "se",
        ys[0]: ses[0],
        ys[1]: ses[1],
        ys[2]: ses[2],
        ys[3]: ses[3],
    })

    reg.append({
        "ctrl": x,
        "type": "p",
        ys[0]: ps[0],
        ys[1]: ps[1],
        ys[2]: ps[2],
        ys[3]: ps[3],
    })
    reg.append(
        {"ctrl": x,
         "type": "n",
         ys[0]: len(df.index)})
```

In [ ]:

In [ ]:

In [12]:

```
regdf = pd.DataFrame(reg)
regdf[ys] = regdf[ys].apply(lambda x: np.round(x, 3))
# regdf.loc[24:, ]
```

In [14]:

```
# regdf
```

In [1]:
```
import numpy as np
import pandas as pd
import os
import regex as re

from collections import Counter
import datetime as dt

import statsmodels.api as sm
from statsmodels.sandbox.regression.gmm import IV2SLS
```

Reanalysis

Missing MSAs

In [2]:
```
# loading ananat's maindata
aej = pd.read_csv("./data/ananat/aej_maindata.csv")
```

In [3]:
```
# loading our version of ananat's maindata
caej = pd.read_csv("./data/replication/casey_maindata_v3.csv")
```

In [4]:
```
# msas ananat had but we didn't have
missing_msas = [i for i in aej.name.values.tolist() if i not in caej.msa.value
s.tolist()]
missing_msas
```

Out[4]:
```
['beaverpa',
 'burlinvt',
 'elmirany',
 'fitchbma',
 'glensfny',
 'grandfnd',
 'iowaciia',
 'kankakil',
 'lawtonok',
 'middlect',
 'muskegmi',
 'norwalct',
 'pittsfma',
 'portlame',
 'portsmnh',
 'poughkny',
 'steubeoh']
```

In [5]:
```
len([i for i in aej.name.values.tolist() if i not in caej.msa.values.tolist
()])
```

Out[5]:
```
17
```

Reanalysis on Ananat's data

In [6]:
```
df = aej.loc[~aej.name.isin(missing_msas)]
df.head()
```

Out[6]:

	name	herf	lenper	hsdrop_w	hsgrad_w	somecoll_w	collgrad_w	hsdrop_b	hsgra
0	akronoh	0.832052	0.001176	0.128893	0.336609	0.269129	0.265369	0.246627	0.35
1	albanyny	0.667636	0.000471	0.134940	0.306759	0.274345	0.283956	0.273940	0.28
2	altoonpa	0.726499	0.000804	0.161442	0.501574	0.199191	0.137794	0.212283	0.48
3	anaheica	0.671304	0.000372	0.139654	0.180415	0.346149	0.333782	0.119261	0.17
4	annarbmi	0.798910	0.000484	0.075142	0.171616	0.262345	0.490897	0.159372	0.23

5 rows × 65 columns

In [7]:
```
df = aej
gini_cols = [c for c in df.columns if re.findall("lngini", c)]
pov_cols = [c for c in df.columns if re.findall("povrate", c)]
ln_cols = [c for c in df.columns if re.findall("ln", c) and c not in gini_cols
]
outcomes = gini_cols + pov_cols + ln_cols
outcomes
```

Out[7]:
```
['lngini_w',
 'lngini_b',
 'povrate_w',
 'povrate_b',
 'ln90w90b',
 'ln10w10b',
 'ln90w10b',
 'ln90b10w']
```

Ananat's complete data

In [8]:

```
df = aej
reg_table = []

for y in outcomes:

    # OLS
    Y = df[y]
    X = df["dism1990"]
    X = sm.add_constant(X)
    mod_ols = sm.OLS(Y, X, cov_type = "HC3")
    res_ols = mod_ols.fit()

    # 2SLS
    df_const = sm.add_constant(df, has_constant='add')
    mod_tsls = IV2SLS(
        endog = df_const[y],
        exog = df_const[["dism1990", "lenper", "const"]],
        instrument = df_const[["herf", "lenper", "const"]])
    res_tsls = mod_tsls.fit()

    coef = {
        "outcome": y,
        "stat": "coef",
        "ols_value": res_ols.params[1],
        "tsls_value": res_tsls.params[0]
    }
    se = {
        "outcome": y,
        "stat": "se",
        "ols_value": res_ols.bse[1],
        "tsls_value": res_tsls.bse[0]
    }
    p = {
        "outcome": y,
        "stat": "p",
        "ols_value": res_ols.pvalues[1],
        "tsls_value": res_tsls.pvalues[0]
    }
    reg_table.append(coef)
    reg_table.append(se)
    reg_table.append(p)

pd.DataFrame(reg_table)
```

```
C:\Users\licas\anaconda3\lib\site-packages\pandas\core\series.py:679: Runtime
Warning: invalid value encountered in reduce
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

Out[8]:

	outcome	stat	ols_value	tsls_value
0	lngini_w	coef	-0.079402	-0.334462
1	lngini_w	se	0.033639	0.113086
2	lngini_w	p	0.019879	0.003746
3	lngini_b	coef	0.459484	0.875067
4	lngini_b	se	0.102096	0.302255
5	lngini_b	p	0.000016	0.004517
6	povrate_w	coef	-0.072789	-0.195749
7	povrate_w	se	0.022422	0.069735
8	povrate_w	p	0.001519	0.005851
9	povrate_b	coef	0.181778	0.258390
10	povrate_b	se	0.051392	0.143716
11	povrate_b	p	0.000578	0.074746
12	ln90w90b	coef	0.111120	-0.130846
13	ln90w90b	se	0.101014	0.287319
14	ln90w90b	p	0.273533	0.649655
15	ln10w10b	coef	1.295175	2.726896
16	ln10w10b	se	0.272151	0.835588
17	ln10w10b	p	0.000006	0.001440
18	ln90w10b	coef	1.171854	1.788736
19	ln90w10b	se	0.285670	0.808050
20	ln90w10b	p	0.000075	0.028776
21	ln90b10w	coef	-0.234441	-0.807314
22	ln90b10w	se	0.132005	0.393780
23	ln90b10w	p	0.078288	0.042564

Ananat's data with 17 cities missing

In [9]:

```
df = aej.loc[~aej.name.isin(missing_msas)]
df.shape
```

Out[9]:

```
(104, 65)
```

In [10]:

```
reg_table = []

for y in outcomes:

    # OLS
    Y = df[y]
    X = df["dism1990"]
    X = sm.add_constant(X)
    mod_ols = sm.OLS(Y, X, cov_type = "HC3")
    res_ols = mod_ols.fit()

    # 2SLS
    df_const = sm.add_constant(df, has_constant='add')
    mod_tsls = IV2SLS(
        endog = df_const[y],
        exog = df_const[["dism1990", "lenper", "const"]],
        instrument = df_const[["herf", "lenper", "const"]])
    res_tsls = mod_tsls.fit()

    coef = {
        "outcome": y,
        "stat": "coef",
        "ols_value": res_ols.params[1],
        "tsls_value": res_tsls.params[0]
    }
    se = {
        "outcome": y,
        "stat": "se",
        "ols_value": res_ols.bse[1],
        "tsls_value": res_tsls.bse[0]
    }
    p = {
        "outcome": y,
        "stat": "p",
        "ols_value": res_ols.pvalues[1],
        "tsls_value": res_tsls.pvalues[0]
    }
    reg_table.append(coef)
    reg_table.append(se)
    reg_table.append(p)

pd.DataFrame(reg_table)
```

```
C:\Users\licas\anaconda3\lib\site-packages\pandas\core\series.py:679: Runtime
Warning: invalid value encountered in reduce
  result = getattr(ufunc, method)(*inputs, **kwargs)
```

Out[10]:

	outcome	stat	ols_value	tsls_value
0	lngini_w	coef	-0.091062	-0.342522
1	lngini_w	se	0.036920	0.115471
2	lngini_w	p	0.015313	0.003762
3	lngini_b	coef	0.494060	0.877026
4	lngini_b	se	0.097263	0.271891
5	lngini_b	p	0.000002	0.001694
6	povrate_w	coef	-0.077727	-0.197244
7	povrate_w	se	0.024983	0.072134
8	povrate_w	p	0.002417	0.007381
9	povrate_b	coef	0.189147	0.245658
10	povrate_b	se	0.055329	0.144705
11	povrate_b	p	0.000906	0.092652
12	ln90w90b	coef	0.074514	-0.161039
13	ln90w90b	se	0.105679	0.282216
14	ln90w90b	p	0.482360	0.569522
15	ln10w10b	coef	1.392569	2.812933
16	ln10w10b	se	0.290872	0.838837
17	ln10w10b	p	0.000006	0.001125
18	ln90w10b	coef	1.257662	1.882302
19	ln90w10b	se	0.304115	0.808456
20	ln90w10b	p	0.000073	0.021891
21	ln90b10w	coef	-0.209421	-0.769591
22	ln90b10w	se	0.144815	0.403833
23	ln90b10w	p	0.151208	0.059530

Reanalysis of our data with 17 missing cities

In [11]:

```
gini_cols = [c for c in caej.columns if re.findall("gini", c) and re.findall(
"ln", c)]
ln_cols = [c for c in caej.columns if re.findall("ln", c) and c not in gini_co
ls]
pov_cols = [c for c in caej.columns if re.findall("pov", c)]
outcomes = gini_cols + pov_cols + ln_cols
outcomes
```

Out[11]:

```
['lnwgini',
 'lnbgini',
 'povrate_w',
 'povrate_b',
 'lnw90b90',
 'lnw10b10',
 'lnw90b10',
 'lnb90w10']
```

In [12]:

```
aej.columns
```

Out[12]:

```
Index(['name', 'herf', 'lenper', 'hsdrop_w', 'hsgrad_w', 'somecoll_w',
       'collgrad_w', 'hsdrop_b', 'hsgrad_b', 'somecoll_b', 'collgrad_b',
       'povrate_w', 'povrate_b', 'mt1proom_w', 'mt1proom_b', 'medgrent_w',
       'medgrent_b', 'medgrentpinc_w', 'medgrentpinc_b', 'area1910', 'passp
c',
       'ngov62', 'manshr', 'incseg', 'closeness', 'regdum1', 'regdum2',
       'regdum3', 'regdum4', 'ethseg10', 'ethiso10', 'ethexp10', 'count1910',
       'black1910', 'ctyliterate1920', 'lfp1920', 'ctytrade_wkrs1920',
       'ctymanuf_wkrs1920', 'ctyrail_wkrs1920', 'count1920', 'black1920',
       'gini_w', 'gini_b', 'p10_w', 'p50_w', 'p90_w', 'p10_b', 'p50_b',
       'p90_b', 'grsrnt_w', 'grsrnt_b', 'lngini_w', 'lngini_b', 'herfscore',
       'ln90w90b', 'ln10w10b', 'ln90w10b', 'ln90b10w', 'mv_st_winus_w',
       'mv_st_winus_b', 'lfp_w', 'lfp_b', 'dism1990', 'pop1990', 'pctbk199
0'],
      dtype='object')
```

In [13]:

```
# slightly renaming ananat's columns to match our version of the data
aej.columns = [f"ln{c[4]}{c[2:4]}{c[-1]}{c[-3:-1]}" if re.findall("ln", c) and
not re.findall("gini", c)
               else c
               for c in aej.columns]
aej.columns = [f"ln{c[-1]}gini" if re.findall("ln", c) and re.findall("gini",
c)
               else c
               for c in aej.columns]
```

In [14]:

```
df = caej[outcomes + ["dism1990", "herf", "lenper"]].append(
    aej.loc[aej.name.isin(missing_msas)][outcomes + ["dism1990","herf", "lenpe
r"]],
    ignore_index = True)
df.shape
```

Out[14]:

```
(121, 11)
```

In [15]:

```
reg_table = []

for y in outcomes:

    # OLS
    Y = df[y]
    X = df["dism1990"]
    X = sm.add_constant(X)
    mod_ols = sm.OLS(Y, X, cov_type = "HC3")
    res_ols = mod_ols.fit()

    # 2SLS
    df_const = sm.add_constant(df, has_constant='add')
    mod_tsls = IV2SLS(
        endog = df_const[y],
        exog = df_const[["dism1990", "lenper", "const"]],
        instrument = df_const[["herf", "lenper", "const"]])
    res_tsls = mod_tsls.fit()

    coef = {
        "outcome": y,
        "stat": "coef",
        "ols_value": res_ols.params[1],
        "tsls_value": res_tsls.params[0]
    }
    se = {
        "outcome": y,
        "stat": "se",
        "ols_value": res_ols.bse[1],
        "tsls_value": res_tsls.bse[0]
    }
    p = {
        "outcome": y,
        "stat": "p",
        "ols_value": res_ols.pvalues[1],
        "tsls_value": res_tsls.pvalues[0]
    }
    reg_table.append(coef)
    reg_table.append(se)
    reg_table.append(p)

pd.DataFrame(reg_table)
```

Out[15]:

	outcome	stat	ols_value	tsls_value
0	lnwgini	coef	-0.155148	-0.103228
1	lnwgini	se	0.090575	0.247665
2	lnwgini	p	0.089332	0.677579
3	lnbgini	coef	-0.346617	-0.046273
4	lnbgini	se	0.131974	0.365154
5	lnbgini	p	0.009764	0.899376
6	povrate_w	coef	-0.087489	-0.108887
7	povrate_w	se	0.021526	0.059030
8	povrate_w	p	0.000087	0.067602
9	povrate_b	coef	0.255816	0.401363
10	povrate_b	se	0.051178	0.143464
11	povrate_b	p	0.000002	0.006013
12	lnw90b90	coef	0.212685	-0.211768
13	lnw90b90	se	0.103509	0.302044
14	lnw90b90	p	0.042091	0.484611
15	lnw10b10	coef	0.432356	1.222478
16	lnw10b10	se	0.281527	0.784555
17	lnw10b10	p	0.127254	0.121869
18	lnw90b10	coef	0.449597	1.593483
19	lnw90b10	se	0.348308	0.983551
20	lnw90b10	p	0.199275	0.107873
21	lnb90w10	coef	-0.195444	0.582773
22	lnb90w10	se	0.184772	0.540589
23	lnb90w10	p	0.292309	0.283218