Potential Differential Undercount in 2020 Census Redistricting Data:

Los Angeles County, California

Paul Ong and Jonathan Ong[1]
UCLA Luskin School of Public Affairs and UCLA Center for Neighborhood Knowledge
August 18, 2021

This Factsheet summarizes the findings from a comparison of population counts for Los Angeles County from the 2020 data for political redistricting (P.L. 94-171 Redistricting Data or PL94) and the 2015-19 American Community Survey (ACS). The Census Bureau conducts an enumeration of the population every decade and compiles the information to assist local officials to redraw political boundaries in response to population changes to ensure that electoral districts are equal in population size. While the goal for every decennial census is a complete and accurate count, it has never been perfect, both missing some individuals and double counting others.[2] One serious problem with miscounting is a differential undercount, where the enumeration systematically undercounts some populations and overcounts other populations. That is, the inaccuracies are not proportionately the same across groups. This problem has profound implications within the redistricting process, essentially disenfranchising those missed by the census and undermining the "one person, one vote" principle. There are also economic consequences because governmental allocation formulas are based on population. Differential undercount is deeply embedded in and shaped by existing structures of inequality. It is, therefore, not surprising that historically low-income persons and people of color are disproportionately missed by the enumeration, thus disproportionately undercounted.[3]

---

[1] Paul Ong is a Research Professor at UCLA School of Public Affairs. Jonathan Ong is a researcher at Ong and Associates, a public-interest consulting firm. Affiliations are for identification purpose only, and authors are solely responsible for the content. We thank Randall Akee and Sonja Diaz for their input, and Chhandara Pech and Melany De La Cruz for their assistance.

[2] It is possible for the overall count to be close to the true population when the two types of errors offset each other.

[3] U.S. Census Bureau, "Census Bureau Releases Estimates of Undercount and Overcount in the 2010 Census," May 22, 2012, https://www.census.gov/newsroom/releases/archives/2010_census/cb12-95.html

To test whether a differential undercount exists for the 2020 census, we develop a new technique to compare the counts to an alternative data source. The U.S. Census Bureau conducts the continuous ACS to gather information on the population and housing, and the Bureau publishes statistics at the census tract level by pooling five years of responses. The results are consistent with the hypothesis that the 2020 enumeration (the basis for PL94) suffers from a differential undercount along economic and demographic lines.

There is circumstantial evidence of a differential undercount for the 2020 enumeration, and the disparities are likely to be considerably larger than in previous decades. This issue is rooted in unforeseen events. The COVID-19 pandemic severely disrupted the 2020 census, creating unforeseen challenges and hurdles not experienced in previous decades.[4] The severity of last year's disruptions forced the Bureau to revise the collection process and extend their timeline.[5] The negative pandemic impacts on the decennial count were unevenly and not randomly distributed. The problem is not just limited to the pandemic. The Trump administration's controversial and politically motivated push to include a citizenship question on the questionnaire further complicated the enumeration.[6] Although the effort was unsuccessful, it nonetheless created fear among immigrants, both legal and undocumented.

Our previous analyses of 2020 self-response rates (those filling out the questionnaire during the first phase of the enumeration) found that the rates to be significantly lower in poor and minority neighborhoods. (See list of references at the end of the Factsheet.) The 2020 self-response rates in disadvantaged communities were also noticeably worse than the self-response rates for 2010. The pandemic disruptions and problems with self-reporting placed greater pressure on the subsequent major stage, on-the-ground outreach and collection efforts for non-responding households. It is unknown at this time if this follow-up drive was successful or not, and the post-enumeration assessment will not be conducted until next year. Other methods and data could be used to determine if there is a differential undercount in the 2020 redistricting numbers.

This Factsheet summarizes one such assessment, a comparison of the 2020 PL94 data against 2015-19 ACS estimates for census tracts in Los Angeles County. A distinct advantage of ACS data is that the collection of responses for those five years were not adversely affected by pandemic, and probably less affected by the political controversy. Interestingly, the two sources have very

[4] The unprecedented problem of conducting collecting responses during the pandemic has also affected other surveys such as the Current Population Survey, Household Pulse Survey and American Community Survey. See list of references at the end of the Factsheet for additional information.

[5] U.S. Census Bureau, "2020 Census Operational Adjustments Due to COVID-19," Last Revised: August 5, 2021, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/operational-adjustments.html

[6] Michael Wines, "2020 Census Won't Have Citizenship Question as Trump Administration Drops Effort," *New York Times*, July 2, 2019, https://www.nytimes.com/2019/07/02/us/trump-census-citizenship-question.html.

similar numbers of reported persons for the whole county, with the PL94 count being 10.01 million and the ACS estimate being 10.08 million. The fundamental analytical question is how those two populations are geographically distributed. This requires using common geographic units. PL94 data are reported in 2020 tracts, while ACS data are reported in 2010 tracts.[7] We reconcile the geographic differences by allocating 2020 counts into 2010 census tracts using a crosswalk provided by the Census Bureau. We were able to merge almost all of the reassigned PL94 tracts with ACS tracts with ACS, accounting for about 99.6% of the 2020 population.

We then examine the relative inconsistency between the two sources. (See Appendix for discussion on the underlying mathematical and statistical concept.) This is done by calculating the percentage difference between the two numbers (PL94 population divided by the ACS population estimates, minus one). A value of zero indicates identical populations, a negative value indicates PL94 is lower, and a positive value indicates PL94 is higher. The results show far less than a perfect match. Two-fifths of the tracts had values less than -2.9%, and another two-fifths had values more than 1.2%. At the extreme ends, one-tenth are less than -10.6%, and another tenth more than 11.5%. These inconsistencies by themselves, however, do not mean that the 2020 enumeration has undercounts and overcounts.

Inconsistencies could be due to the ACS being a survey that samples only about 2%-3% of all households per year (five times that when using 5-year statistics). We need to determine if the differences are due simply to sampling error, or if the discrepancies are larger than reasonably expected. If the latter, then we examine if the differences are systematic along economic and demographic lines. The Census Bureau reports a margin of error for each tract, which defines the range where there is a 90% chance that the "real" population count would fall. In other words, we would expect approximately 5% to be higher than the range and approximately 5% to be lower than the range. If we hypothetically assume that the inconsistency between PL94 and ACS to due solely to sampling error, then we would expect that roughly 5% of the tracts to have percentage differences above the corresponding range, and roughly 5% of the tracts to have percentages differences below the corresponding range.[8] The findings do not support these hypotheses: about 13% of tracts are above the 90% confidence interval, and over 16% of the tracts are below. This indicates that there are places where the PL94 undercounts relative to the ACS, and other places where the PL94 overcounts relative to the ACS estimates.

The next step is to determine if the discrepancies are systematic rather than random. Non-random patterns are necessary albeit not sufficient conditions for differential undercount. We test for systematic differences by examining how the percentage difference between the two numbers (PL94 population divided by the ACS population estimates, minus one) vary along

[7] There were minor changes to the 2010 census tracts made in 2012, which are accounted for in the crosswalk.

[8] The percent is approximate because the 5% figure is hypothetical with an infinite number of observations. Although the dataset has a large number of observations (over 2,000), there is still a random chance for slightly higher or lower numbers falling outside the range.

socioeconomic and demographic lines, particularly those found to have contributed to a differential undercount in the past. Tract characteristics are based on the 2015-19 ACS. Figures 1-4 presents our findings. A positive value indicates that the PL94 overcounts relative to the ACS, and a negative value indicates the PL94 overcounts relative to the ACS estimates. Another way to interpret the graphs is the size of the vertical gap between two bars, with a sizable gap indicating a differential undercount between the two groups.

Figure 1 reports the differences by neighborhoods defined by the race/ethnic group that comprise more than 66% of the population within a tract. All other tracts are categorized as "Other". The results reveal a systematic racial/ethnic variation. Non-Hispanic White neighborhoods have positive differences, and Hispanic neighborhoods have the lowest, a difference 4.4 percentage points. This pattern is consistent with previous post-enumeration studies prior to 2020.

Figure 1:

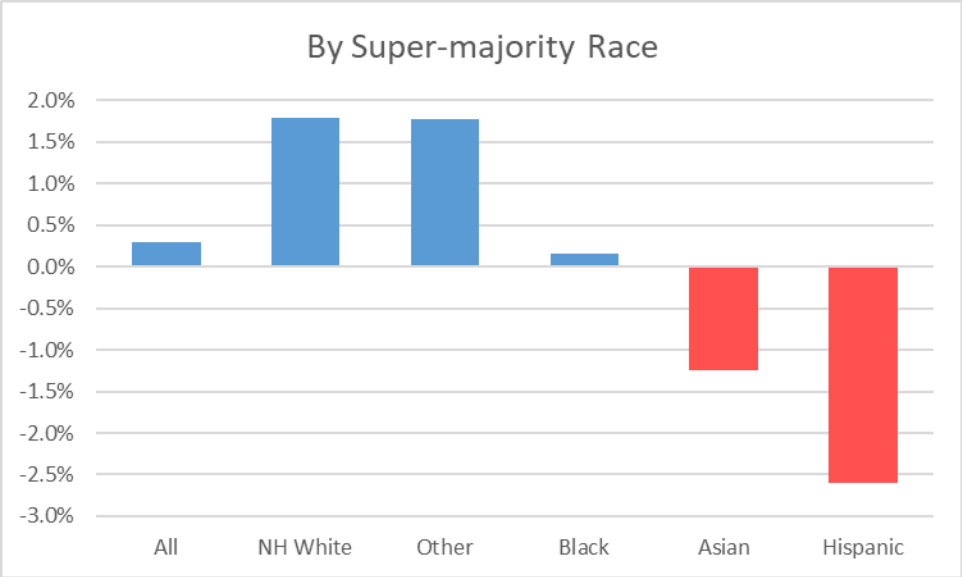


Figure 2 reports the differences by neighborhoods defined by the percent of households that are renters. The categories are created by dividing the tracts into quintiles (five groups of equal size), ranging from the group with the highest percentage of renters to the group with the lowest percentage. The results show that tracts composed predominantly of homeowners have higher positive differences. Neighborhoods that are disproportionately renters tend to have lower differences, although the pattern is not linear. The tracts with the highest proportion of renters have a positive difference, although not as high as for neighborhoods with mostly homeowners. This outcome may be due to some particular unobserved socioeconomic characteristics. The largest gap is between the lowest renter quintile and the second highest renter quintile, about 2.5 percentage points. The findings are generally consistent with previous post-enumeration studies for 2010 and prior earlier decades.

Figure 2:

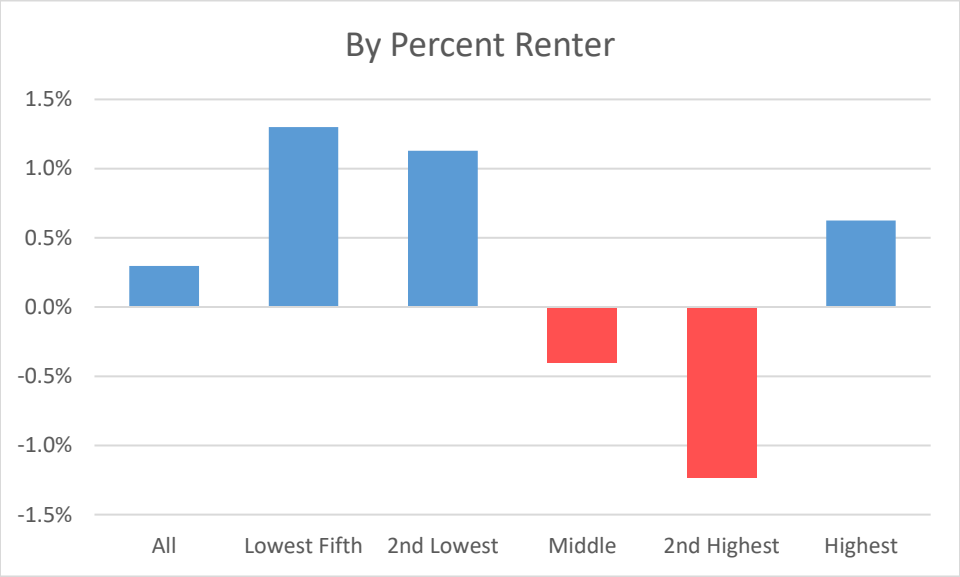


Figure 3 reports the differences by neighborhoods defined by the percent of individuals falling below the federal poverty line. The categories are created by dividing the tracts into quintiles, ranging from the group with the lowest percent in poverty to the group with the highest. The results generate a consistent linear pattern. The least poor neighborhoods have the highest positive differences, while the poorest places have the most negative differences. The difference between the groups at the two extremes (the lowest and highest quintiles) is about 4.3 percentage points. The findings are consistent with previous post-enumeration studies prior to 2020.

Figure 3:

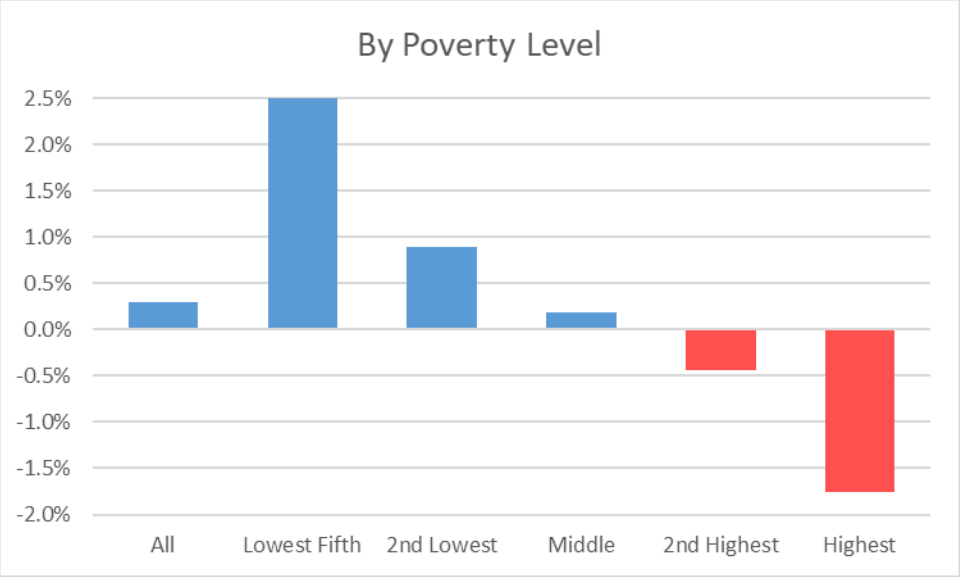
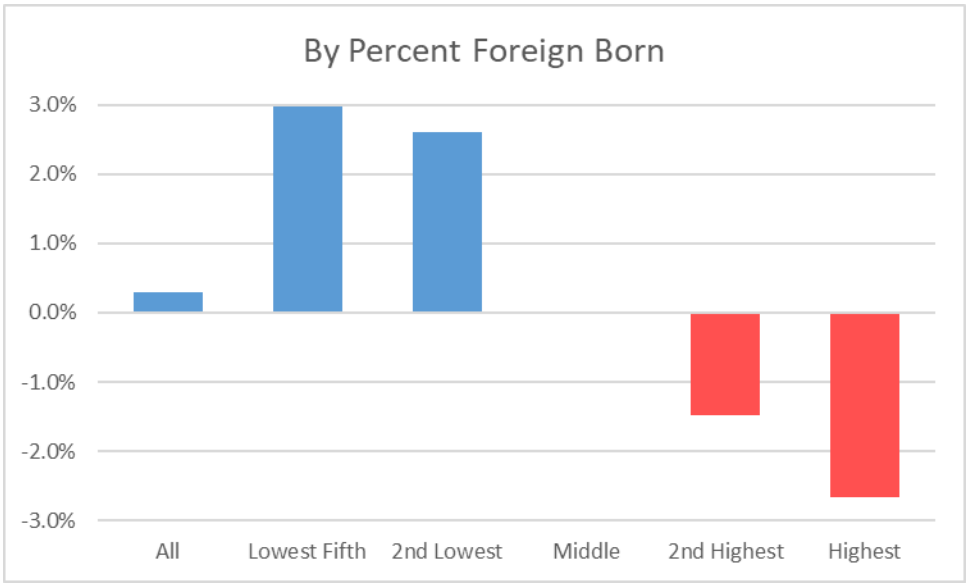


Figure 4 reports the differences by neighborhoods defined by the percent of individuals who are foreign born. The categories are created by dividing the tracts into quintiles, ranging from the group with proportionately the least number of foreign-born individuals to the group with proportionately the most. The results show a consistent linear pattern. Neighborhoods that are mostly U.S. born (the complement of foreign born) have highest positive differences, while the places with the most number of foreign born have the most negative differences. The difference between the groups at the two extremes (the lowest and highest quintiles) is about 5.6 percentage points. This spread is the greatest among the four comparisons (race/ethnicity, renter/owner, poverty level and nativity), suggesting that the controversy around the citizenship questions contributed to an undercount. Like the other three analyses, the findings in Figure 4 are consistent with previous post-enumeration studies.

Figure 4:



We conducted additional analyses to test the robustness of the above findings. This includes restricting the sample of tracts to those that have a one-to-one match between ACS and PL94 (that is, no changes in boundaries) and deleting outliers. To account for collinearity among the factors (e.g., poverty and race are correlated), we ran several multivariate regressions, which also included other variables that are potentially associated with low response rates (e.g., median household income, percent residing in group quarters, availability of high-speed internet, past growth rate, and percent of the population between the ages of 0 to 4). The findings from these analyses are consistent with the patterns reported in Figures 1 to 4.

The overall conclusion is that the available data (along with previous studies and other circumstantial evidence) demonstrates that a differential undercount is very likely. In other

words, race, poverty, being a renter and being foreign born matter in determining who is included or excluded from the 2020 enumeration.[9]

While our analysis is a concrete test for differential undercount, it has limitations. The most important is the assumption that the ACS is a useful benchmark because its data collection is not adversely affected by the COVID-19 pandemic and the politicization of the questionnaire. This seems reasonable. This assumption is supported by a comparison of the 2010 census and the 2006-10 American Community survey.[10] That comparison finds that the differences between the previous 2010 enumeration counts and the ACS population estimates are primarily associated with sampling error. This is very different than the results of the comparison of the 2020 census and the most recent ACS population estimates, which are presented earlier in the Factsheet. The discrepancy between the two decades is additional evidence that the COVID-19 pandemic adversely impacted the coverage and fairness of the data collection for the 2020 enumeration, and that the 2020 counts are less accurate.

Finally, it is possible that the ACS also has systematic biases.[11] Those biases, however, are more likely to lead to differential underestimates in ACS, thus attenuating the gap between ACS and PL94. In other words, our method may understate the magnitude of the differential undercount in the 2020 enumeration.[12]

Clearly, additional research will be needed to further test for the existence of systematic differential undercount, and equally as important, to determine the magnitude of the variations. This is critical because census numbers have profound political, economic and social consequences. It is vital that we have an accurate portrait of America that is inclusive and fair.

---

[9] The technique used in our analysis is innovative, and is a potential tool for adjusting differential undercount. Previous techniques rely on imputing results from the post-enumeration studies to simulate the undercount for local neighborhoods. This approach assumes that national-level findings from a small representative number of case studies are the same small geographies everywhere. Moreover, it assumes that the differential undercount patterns from the previous decade are the same for the current decade. Although these assumptions may be reasonable, there is no empirical evidence to support the underlying assertion. Our technique, on the other hand, relies on data specific to each local neighborhood, and is therefore potentially more accurate and precise.

[10] Ideally, we prefer utilizing 2005-09 ACS population estimates, but no such dataset is available. When using the available 2006-10 ACS as a benchmark, we find that approximately 13% of the tracts had 2010 census counts that were outside of the confidence interval for the ACS estimates, which is slightly larger than the expected number based on the margin of error. This compares to 30% of the tracts when comparing 2020 census counts and the most recent ACS estimates.

[11] Another potential source of potential error comes from the fact that the ACS is partially dependent on population estimates, which were last benchmarked against the 2010 census. U.S. Census Bureau, "Population and Housing Unit Estimates," https://www.census.gov/programs-surveys/popest.html.

[12] Consider the following scenario. ACS undercounts a neighborhood by 5%, and PL94 is 10% lower than ACS estimate; therefore, the total undercount is nearly 15%.

==========

References:

Akee, Randall; Ong, Paul M; Rodriguez-Lonebear, Desi. “US Census Response Rates on American Indian Reservations in the 2020 Census and in the 2010 Census.” Technical Report. UCLA Center for Neighborhood Knowledge and American Indian Studies Center.

Ong, Paul M and Ong, Jonathan. August 18, 2020. “Persistent Shortfalls and Racial/Class Disparities”. Technical Report, UCLA Asian American Studies Center, UCLA Center for Neighborhood Knowledge, and Ong & Associates, 2020.

Ong, Paul M and Ong, Jonathan. June 11, 2020. “Persistent Shortfall and Racial/Class Disparities, 2020 Census Self-Response Rate.” UCLA Luskin School of Public Affairs and UCLA Center for Neighborhood Knowledge. UCLA Latino Policy & Politics Initiative and Center for Neighborhood Knowledge.

Ong, Paul M; Ong, Elena; and Ong, Jonathan. May 7 and May 12, 2020 “Los Angeles County 2020 Census Response Rate Falling Behind 11 Percentage Points and a Third of a Million Lower than 2010,” Technical Report. UCLA Center for Neighborhood Knowledge.

APPENDIX: MATHEMATICAL DETAILS OF ANALYSIS OF DIFFERENTIAL BIASES IN PL94 COUNTS

This appendix provides the underlying mathematical concept behind the analyses of possible differential undercount.

Basic Equations

First, consider the following two equations for ACS and PL94 (PL for short):

$$(1)\quad ACS_i = P_i + f(X_i) + \varepsilon_i \text{ , for tracts = 1… n}$$

ACS is the estimated population, P is the actual population, and ε is a normally distributed random term with expected value of zero. The term ε is a product of sampling error. The function f(X) captures any potential systematic biases in ACS due to vector of X variables based on the literature and previous research. The second equation is:

$$(2)\quad PL_i = P_i + g(Z_i) + \mu_i \text{ , for tracts = 1… n}$$

PL is the population count from PL94, P is the actual population, and μ is a random term with expected value of zero. The function g(Z) captures systematic biases in PL due to vector of Z variables based on the literature and previous research. In other words, the function captures differential undercount. Subtracting (1) from (2) yields:

$$(3)\quad PL_i - ACS_i = g(Z_i) + \mu_I - f(X_i) - \varepsilon_i \text{ , for tracts = 1… n}$$

Equation (3) is then transformed by dividing by ACS

$$(4)\quad (PL_i / ACS_i) - 1 = [g(Z_i) + \mu_I - f(X_i) - \varepsilon_i] / ACS_i \text{ , for tracts = 1… n}$$

The term on the left is the percent difference between PL and ACS. If we assume

$$(5)\quad D_i = (PL_i / ACS_i) - 1 \text{ , for tracts = 1… n}$$

D can be interpreted as evidence of possible differential undercount. A positive value implies a relatively higher count compared to other groups (over count), and a negative value implies a relatively lower count compared to other groups (under count).

Testing ACS Sampling Error

If we assume that (a) PL is complete, accurate and unbiased enumeration, and (b) ACS is unbiased, then we are left with.

$$(6)\quad D_i = -\varepsilon_i / ACS_i \text{ , for tracts = 1… n}$$

We test this assumption of whether the inconsistency between PL94 and ACS is due to sampling error. We conduct this test by using the Census Bureau's margin of error for each tract, which we then can construct the 90% confidence interval around each D. For each tract, we can determine whether D falls within that range. If assumption for (6) is true, then we expect roughly 5% to be above the range and 5% to be below the range. The analysis shows that significantly more observations are outside the range, thus (6) is not true.

Testing PL94 for Systematic Biases

We can now assume that (a) PL94 is not "complete, accurate and unbiased", and (b) ACS is not biased. We can rewrite (4) as:

$$(7)\quad D_i = g(Z_i) + \delta_i \text{ , for tracts = 1… n}$$

The last term, δ, is the combined random error term, and the functions g(Z) is divided by ACS. We assume a linear form, which can be tested using OLS regressions.

(8) $D_i = \alpha + \beta * Z_i + \delta_i$, for tracts = 1… n

Both β and Z are vectors. If there is systematic bias, then the estimated coefficients would be statistically significant. The results show that the estimated coefficients are highly significant for race, nativity, income, poverty, past growth rate, and other factors. This is true for both the entire sample of tracts, and for a more restricted sample of tracts (tracts with identical geographic boundaries in ACS and PL94, and excluding outlier). We also ran the following alternative specification:

(9) $PL94_i = ACS_i + h(Z_i) + \eta_i$, for tracts = 1… n

We assume that the function h(Z) has a first-order linear form (vector $\lambda * Z_i$) and estimate the coefficients using OLS regressions. We get the same qualitative results.

Possible ACS Systematic Biases

The empirical results from the above have limitations due to the assumption. If ACS also has systematic biases, then equation (7) would include another term:

(10) $D_i = g(Z_i) - f(X_i) + \delta_i$, for tracts = 1… n

Unfortunately, we do not have enough information to decompose the two potential sources of systematic biases in D. We are, however, able to offer reasonable conjectures on the likely effects on the estimated coefficients for (8). We can separate the potential biases into those that pre-date the pandemic and persisted into the pandemic. For example, this overlap is likely to be true for race, nativity, and poverty. Call the overlapping causes vector V. For any given variable in this subset, it is likely that the direction of the effects is in the same direction for the first derivative:

(11) If $d[g(v)]/dv > 0$ then $d[f(v)]/dv > 0$ and

(12) If $d[f(v)]/dv < 0$ then $d[f(v)]/dv < 0$.

The implication is that the OLS regressions are underestimating the systematic biases in PL94. For example, if both ACS and PL94 undercount Latinos, then the estimated impact from (8) is a combination of the biases in both data sources. The undercount in PL94 is that estimated value plus the bias in ACS. As mentioned, the second source is unobserved. It is possible that there are also non-overlapping potential causal variables. A primary example is the variation in COVID-19, which is only applicable to the 2020 enumeration. This requires having tract level data, which is currently not available.