Final Replication Project:
Project Report

Sewheat Haile

May 14, 2021

1 Tables and Figure

The values in my tables and figure look very similar to those in Mandel and Semyonov's tables and figure. However, my tables consistently report a larger N than the authors'. This could be due to differences in the sample obtained through IPUMS. This could also be due to differences in cleaning procedures of the data. For example, the authors may have determined a person's sector using the variable IND, whereas I chose to use the variable CLASSWKR. They may have also used EMPSTAT to determine which respondents are economically active, whereas I used LABFORCE.

Although only by negligible amount, the means in my Table A1a are consistently lower than the authors'. The means in my Table A1b are also slightly lower or higher than the authors' in some cases. These discrepancies are likely due to the larger N in my analysis. My indices of dissimilarity are almost identical to the authors' in Table A1a and are quite similar in Table A1b as well.

Most of my regression coefficients in Table A2a are similar or identical to those in the authors' table. However, there are two discrepancies in the coefficients for "has kids under age 5" where I report negative values and the authors report positive values. The coefficients in my Table A2a for 1980 White and 1990 Black are -0.002 and -0.0003 respectively, compared to the authors' 0.004 and 0.008. It is important to note both of my coefficients are not significant, whereas the authors' value for 1980 White is significant (the value for 1980 Black is not significant). Because my values and one of the authors' values are not significant, the differences in sign seem to be unimportant. In fact, the only coefficient in my Table A2a that is significant for "has kids under age 5" is 2000 White, whereas the authors' variable had a total of six significant coefficients. Interestingly, the discrepancies for "has kids under age 5" in Table A2a do not exist in Table A2b; my analysis closely matches the authors'. However, the coefficients in my Table A2b for "less than high school" are slightly larger than the authors' coefficients by about 0.1 to 0.2. The discrepancies in both my Table A2a and Table A2b could be due to differences in IPUMS samples, or due to the larger N in my analysis.

One question I have for the authors is whether they used EDUC or EDUCD to determine a person's years of schooling. These two variables are coded quite differently and may contribute to differences in table numbers if the authors and I did not use the same variable. For example, I cleaned for years of schooling using EDUCD as opposed to EDUC. EDUCD has more categories than EDUC, such as “master’s degree,” “doctoral degree,” and separate categories for grade 1 through 4. These observations are included in EDUC under its less detailed categories, such as ”5+ years of college” and the singular category ”grade 1-4”. If the authors chose to use EDUC instead of EDUCD, that could explain the differences in values for potential work experience (age - years of schooling - 6) between my tables and theirs.

2 Room for Improvement

I understand the authors’ decision to compare only white and black non-Hispanic men and women as additional race categories would create complicated tables and figures. However, restricting the analysis to only these racial categories provides an oversimplified picture of the racial pay gap between men and women, particularly in later years where Hispanics/Latinos make up a larger share of the US population. I would extend the analysis by including two additional ethnoracial categories—Asian and Hispanic/Latino of all races (thus forward referred to as ”Latino”). The analysis for the Latino category may be particularly challenging given its census designation as an ethnicity rather than a race. It is likely that phenotypically black or *mestizo* Latino people receive lower wages than phenotypically white Latinos. Lumping all of these phenotypically different groups into one category will not capture the heterogeneity in wages within the Latino category. However, one could also analyze the distribution of wages within each race category to determine whether there is greater heterogeneity among Latinos than among other racial groups.

I also feel that it was a mistake to exclude the “other” race category in the authors’ original analysis. Approximately 40% of Latinos mark “other” as their race on the census [2], and 97% of the “other” race category is comprised of Latinos [1]. I believe that combining the “other” and “Latino” categories would further strengthen this study by accounting for a greater number of Latinos who would otherwise be excluded from this study.

References

[1] Compton, Elizabeth, Michael Bentley, Sharon Ennis, and Sonya Rastogi. 2012. “2010 Census Race and Hispanic Origin Alternative Questionnaire Experiment.” Washington, DC: U.S. Census Bureau.

[2] Miyawaki, Michael Hajime. 2016. ”Part-Latinos and racial reporting in the census: An issue of question format?” *Sociology of Race and Ethnicity* 2.3: 289-306.