Lab 2: Describing Generations

A Contemporary Examination of Equal Pay across Generations

Hong Hu, Michael Guldberg, Sohail Khan, Joseph Eagan

Contents

1 Importance and Context

Equal Pay across generations is the interested variable we chose to analyze because of the effects generations (age), race, gender, and education have on income distribution.

Our analysis thus focuses on this key question:

> *What is the current status of equal pay across generations?*

This analysis develops preliminary answers about the selected variables effect on income variation across generations. Those selected variables are race and gender on income. Hence, the analysis takes into consideration ageism, The Civil Rights Act of 1964 and The Equal Pay Movement. To operationalize our dependent variable— generations (age), we grouped generations by age (e.g. Baby Boomer, Gen X, Millennial, Gen Z, Gen Alpha). Moreover, we categorized these variables to accurately represent the workforce. We inspected other variables (Marital Status and Number of Siblings) but abandoned them due to time constraints. Importantly, the small effect of the variables deserves further exploration.

2 Data and Methodology

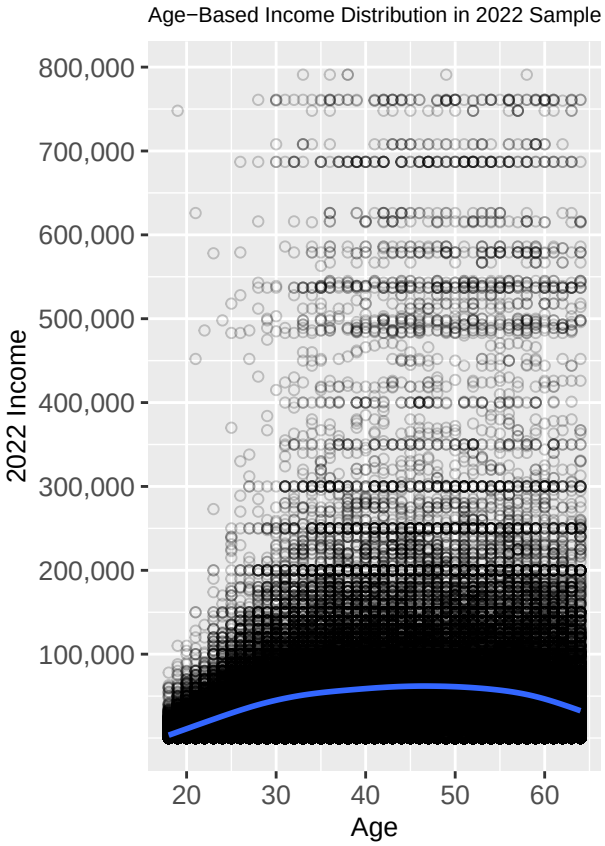
We utilized data from the 2022 American Community Survey, which was provided by IPUMS USA[1]. This dataset consists of a 1-in-1000 national random sample of the population, totaling 338,175 samples spanning all age groups. Our dataset includes variables such as age, gender, and race, which serve directly or indirectly as the input variables, while wage income is designated as the output variable.

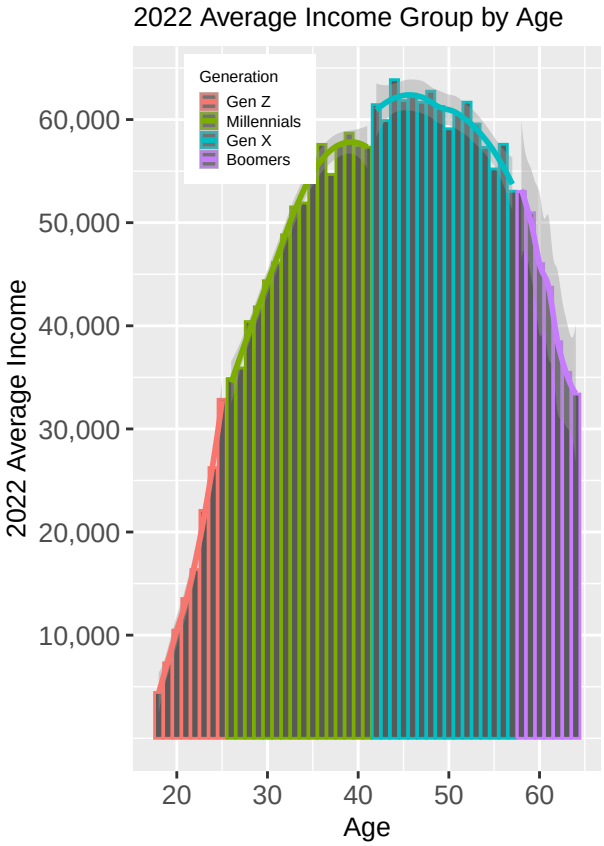
We created a robust and efficient data extraction pipeline from the IPUMS USA Dataset.Given that certain occupations restrict individuals under the age of 18 and the conventional retirement age is 65, we have

[1]IPUMS USA: https://www.ipums.org/

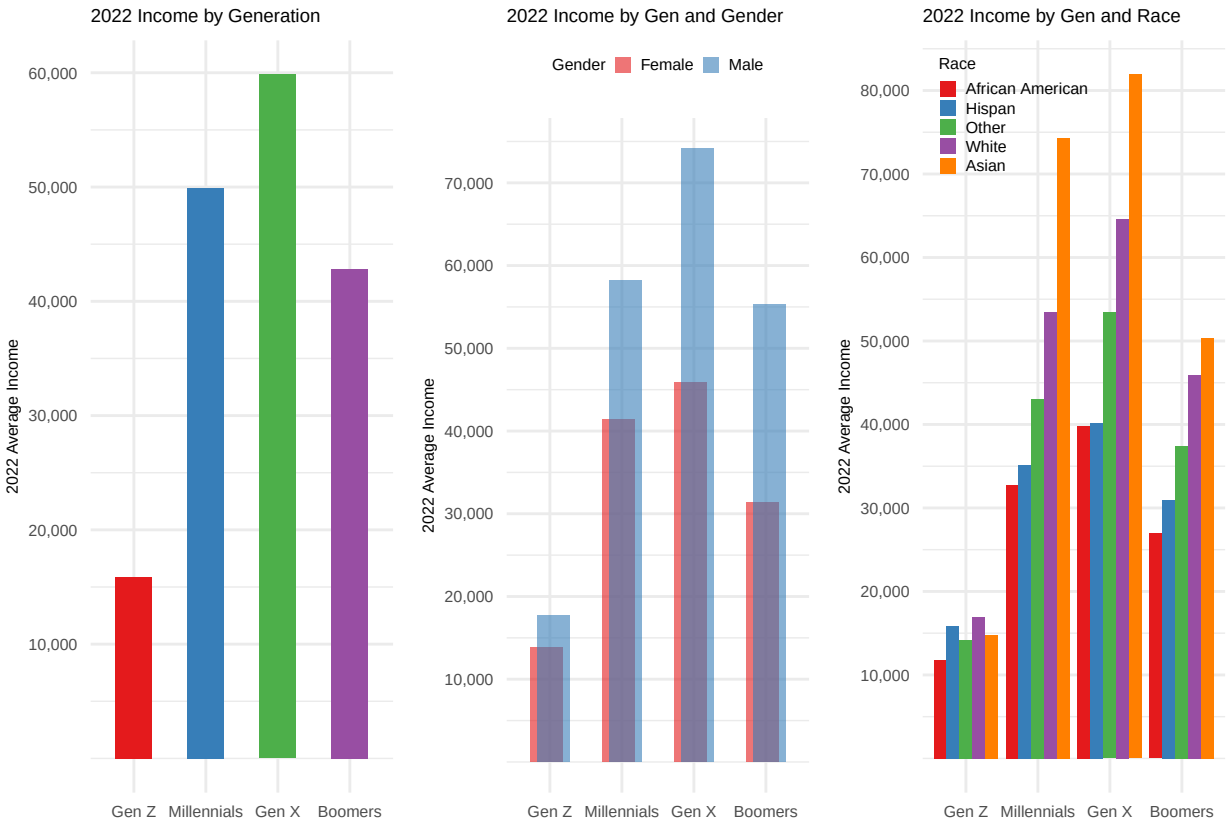
selected wage income data specifically for individuals aged between 18 and 64 from the sample to better demonstrate equal pay across generations. After selecting the subset of the data, we created our own variable to define race category (race_cat). In spirit of comprehension, we then mutated several columns into more understandable and tangible formats. We assigned "Male" and "Female" to our gender columns, and then created a specialized Generation column. This generation column was defined by parsing by a range and group of ranges based on the age range of our data.

Following data wrangling and utilizing 30% of the records for exploration, we analyzed the remaining 138,300 samples using the left point chart displayed below. Observing the smoothing line, we note a quadratic distribution of the sample CEF on mean wage over age. Our later Polynomial Approximation of the CEF also proves it. The accompanying zoom-in bar charts on the right also illustrate a similar trend: average wage income experiences a significant surge within the Gen Z group, followed by a rapid and steady increase among Millennials. It peaks around age 48 within the Gen X group, gradually declining until reaching the retirement age for Baby Boomers. An intriguing observation is the notable increase in wage income beyond $650,000, which approximates the maximum tax bracket threshold (37% tax rate) for the year 2022, with subsequent salary increments averaging around $50,000.





Prior to commencing our regression modeling, we conducted a brief examination using the following bar charts to assess the strength of association between our candidate input variables—generations, gender, and race—and wage income. The visualizations clearly depict a strong correlation between all three categorical variables and wage income, signifying substantial differences in wage income across different races, genders for each generation. Consequently, it is prudent to analyze how these variables, in conjunction with generation groups, are associated with the outcome variable of yearly salary. Additionally, noteworthy findings include relatively equal salary distributions among different races within the Gen Z cohort, possibly due to all members of this group still accumulating work experience with lower salaries compared to other generational cohorts.



For this analysis, we created 4 models with the intent to incrementally understand the relationship between average wage, age/generation, and various other factors by strategically adding covariates. To do this, we first created a baseline model (Model 1) which regresses wage on a transformed age variable. As the baseline model is quadratic, hard to explain its result, we then created three additional models which account for race, gender, and highest completed degree across generations (Models 2, 3, 4). These variables were chosen based on hypothesis about the factors that may be related to wage. Additionally, these models transform the age variable to generation in order to make their results more interpretable.

Before conducting the regression analysis, it's essential to evaluate the underlying assumptions. In the context of our large-sample ordinary least squares (OLS) modeling, the initial assumption to verify is the independence and identical distribution (IID) of the data. Given that IPUMS USA has provided us with a 1-in-1000 national random sample, we can reasonably assume that this condition holds. The second assumption pertains to the presence of a Best Linear Predictor (BLP). While the distribution of income may pose a challenge due to its significant variance, we typically address this issue by applying a logarithmic transformation to the income variable. However, considering that some samples may have zero income, we have opted to include both age^2 and age as input variables in the age over income regression (Model 1) to mitigate this concern. Regarding other input variables of generation, race and gender, given that they are categorical variables, we can assume that the covariance between these variables and the output variable is finite, thus making the existence of a BLP. The final assumption to consider is the uniqueness of the BLP, which entails the absence of perfect collinearity among input variables. It's evident that there's no strong correlation among the input variables. In summary, although the BLP assumption may face challenges, it remains manageable, and we can affirm that all the requisite assumptions for large-sample OLS regression have been satisfied.

Table 1: Relationship Between Income and Generation

	Dependent variable:			
	incwage			
	(1)	(2)	(3)	(4)
age	7,217.371*** (80.059)			
I(age^2)	−79.472*** (1.012)			
generationMillennials		34,059.510*** (337.144)	33,301.380*** (335.688)	34,229.120*** (339.192)
generationGen X		43,960.520*** (428.802)	42,958.900*** (425.584)	44,397.820*** (429.249)
generationBoomers		26,934.990*** (505.152)	25,326.560*** (503.956)	27,656.910*** (504.824)
race_catHispan			2,398.658*** (524.562)	
race_catOther			9,939.520*** (834.232)	
race_catWhite			19,026.740*** (471.084)	
race_catAsian			32,626.940*** (1,006.088)	
genderMale				19,667.810*** (372.837)
Constant	−101,129.400*** (1,339.019)	15,879.830*** (152.943)	1,248.692** (429.160)	5,739.085*** (246.252)
Observations	138,300	138,300	138,300	138,300
R^2	0.052	0.043	0.058	0.062
Adjusted R^2	0.052	0.043	0.058	0.062
Residual Std. Error	69,338.240 (df = 138297)	69,639.960 (df = 138296)	69,094.960 (df = 138292)	68,942.790 (df = 138295)

Note: *p<0.05; **p<0.01; ***p<0.001

3 Results

Our analysis of wages across generations, incorporating gender, race, and educational factors, reveals multiple statistically significant insights. With Generation Z serving as the reference group, Generation X members are shown to have the greatest average wage advantage of approximately \$43,960.52 ($p < 0.001$) over Generation Z (Model 2). In contrast, members of the Millennial and Baby Boomer generations experienced lesser wage advantages over Generation Z, averaging increases of \$34,059.51 ($p < 0.001$) and \$26,934.99 ($p < 0.001$) respectively (Model 2). When examining Model 3, we see significant results across all races, with the reference category being African American. An important observation from these results is that average wages for all races are higher than the African American reference group, with average advantages ranging from \$2,398.66 to \$32,626.94 ($p < 0.001$). When examining gender, we see a significant discrepancy between males and females, with average male wages being \$19,667.81 ($p < 0.001$) more than females (Model 4).

4 Discussion

The results of this analysis are important to understanding the current state of wage equality across generations. There is significant evidence across the board that generation, race, and gender all play important roles when describing wages. Potential explanations for the generation specific findings include a lack of experience for members of Generation Z and the conclusion of careers for member of the Baby Boomer generation. In contrast, Generation X's large wage advantage may underscore their strong position in the job market due to significant experience and valuable positions within their career trajectories. However, it is important to emphasize that while this study's findings are statistically significant, they are not causal, and these potential explanations require further exploration. This analysis also highlights significant differences in wages when considering gender, race, and education as well as generation. These results are alarming, and once again warrant further investigation into potential causal relationships between race and wage and gender and wage. The low R^2 values across all models indicates that these models leave a large proportion of variance in income unaccounted for by X variables. These may include variables such as industry, occupation, work experience, and education, which are excluded from this analysis in order to focus on insights related to equal pay.