Analysis of Effect of Physical Characteristics on Total Personal Income of People with STEM Occupations

[redacted]

12/16/2021

Abstract

The purpose of our analysis is to investigate the relationship between total income of people with STEM occupations and physical characteristics and demographics such as sex, race, and age. In our study, we use data set from 2019 IPUMS USA Census Data for social, economics, and health research. Also, to analyze the effect of sex, age, and race on the total personal income, we will use multiple regression models. Our analysis shows the sex and race affect toatal personal income but there was no significant evidence that age affects total personal income. This result provides the profound evidence of the existance of racial and sexual inequlities in STEM field occupations.

Background and Purpose

We are now living in an era where technology is almost everywhere around us: smartphones, computers, internet games, applications, etc. As the demand for those technology has been continuously increase and our economy hugely depends on the impact of STEM, the STEM occupations are getting more attention lately. Pew Research Center showed that "since 1990, STEM employment has grown 79% (From 9.7 million to 17.3 million), whereas overall employment grew only 34%[1]." Moreover, the U.S. Bureau of Labor Statistics (BLS) even shows that occupations in the STEM field are expected to grow 8.0% by 2029, compared with 3.7% for all occupations[2]. However, at the same time, there still exists a gender and racial inequality in STEM field just like other fields of occupation. For instance, Yonghong Xu, an associate professor in the Department of Counseling, Educational Psychology, and Research at University of Memphis, argues that women in STEM occupations are underrepresented from the aspect of earning differentials. Through the research, she investigated that "women in STEM occupations experienced multiple earning penalties concurrent with their growing family obligations[3]."

In the light of this finding, our project will explore how physical characteristics (sex, race, and age) which construct the part of social identities can affect one's total personal income[4]. Since we are interested in the inequalities in STEM fields, we will focus on the people who work in the STEM fields. In our research, we first will analyze how sex and race can affect one's total personal income. Here, we will create four different groups: white male, white female, POC male, and POC female. Then, we will deepen our research by investigating the effect of age on one's total personal income. Please note that we only look at the total personal income of people who are 18 or older since a person is considered as a legal adult at the age of 18 in the most of the states in the United States; Alaska and Nebraska set it to 19 and Mississippi sets it to 21[5].

Our primary hypothesis is among people who have STEM occupations, white males have higher personal total income compared to women of color, on average. Then, we also hypothesize that among people who have STEM occupations, the difference in personal total income between white males and women of color gets larger as the workers gets older.

Methods

Data

We will use data from IPUMS USA. IPUMS USA is "a database providing access to over sixty integrated, high-precision samples of the American population drawn from sixteen federal censuses, from the American Community Surveys of 2000-present, and from the Puerto Rican Community Surveys of 2005-present". Since we are particularly interested in the current situation in the U.S., so we decided to use the data of 2019. From this database, we extracted information on the subject's sex, age, race, educational attainment, occupation, and total personal income. In our dataset, we also included a detailed categorization of race and educational attainment. Under race, for example, some categories were Okinawan, Mongolian, and several Native American Tribes. For detailed educational attainment, some groups including missing educational attainment, masters, and Ph.D. attainment were included.

Variables

Variables we explicitly included in our analysis of the data consisted of: Sex, Age, Race, Occupation and total personal income. Our response variable was total personal income, and our explanatory variables were age, sex, occupation and race depending on our hypothesis. In order to analyse our regression model accurately, we also excluded all data that had a missing total personal income value. Any ages below 18 were also excluded in order to examine income for potentially higher income levels, or positions held by more experienced workers.

Some variables that we used that were separated into a binary form were sex, and race. Although we had several groups of races, we combined minority groups into one called "POC". All of our data points chose either female or male gender, so we separated it into a binary explanatory variable. In our secondary hypothesis, we chose three ages for each quadrant of our data in order to represent each age range. Early career was represented by all data points with 29 years of age, mid career with 44 years of age, and lastly, late career was represented by 59 years of age. Another variable separated into a binary occupation which was separated into STEM field occupations and non-stem field occupations. Our categorization was based on the 2018-onward census occupational classification system, which indicated the individual categories of occupation varying from business to natural resources. Details of occupation codes can be found here.

Analysis

To analyze the primary hypotheses, we fit two different additive multiple regression models and one interactive multiple regression models.The first model is the simplest additive model which only includes one explanatory variable, SEX. Then, we add a complexity to the second additive model by adding another explanatory variable which is RACE. Lastly, we build the interactive model with SEX and RACE For the secondary hypotheses, we fit two models; one without interaction of AGE and one with interaction of AGE. After finding the right model for each hypothesis, we perform contrasts on the model predictions because we are particularly interested in two groups: white males and women of color. We performed individual z-tests to see the significance of each coefficient, and nested F-tests to compare the overall goodness of fit of models.

Results

By filtering out all non-applicable observations, we were left with the sample size of 119,568 with 86,679 males (72.49%) and 32,889 females (27.51%), with media age of 43 years. Also note that the sample contains 90,078 whites (75.34%) and 29,490 people of color (24.66%). The median personal total income is $76,000.

When we conducted nested F-tests, we used 0.95 or 95% as our cut point. Thus, we reject the null if the p-value is less than 0.05 and conclude that we fail to reject the null if the p-value is greater than 0.05.

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
119566	7.044318e+14	NA	NA	NA	NA
119565	7.041212e+14	1	310652473851	52.75109	0

Nested F-test between Model 1 and Model 2

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
119565	7.041212e+14	NA	NA	NA	NA
119564	7.040830e+14	1	38203591366	6.487551	0.0108645

Nested F-test between Model 2 and Model 3

The nested F-test shows that the additive model with two explanatory variables, SEX and RACE, explains more variability in personal total income (INCTOT) than the simple linear regression model with SEX as an explanatory variable since its p-value is 3.809e-13 which is extremely smaller than 0.05. Thus, we preferred the second model over the first model. We, then, perform another nested F-test between the second model and the third model which allows the interaction between SEX and RACE. Here, we found that the third model explains more variability in personal total income than the second model because the p-value is 0.01086 $>$ 0.05. This leads us to conclude that the interaction term (SEX * RACE) is a necessary component of the model. Moreover, the individual z-tests also show that the coefficients for every explanatory variable is significant to the third model.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	95978.956	296.982	323.181	0.000	95396.876	96561.037
SEX: Female	-25418.819	583.794	-43.541	0.000	-26563.046	-24274.592
RACE: POC	2872.787	619.627	4.636	0.000	1658.328	4087.246
SEX: Female:RACEPOC	2849.274	1118.649	2.547	0.011	656.740	5041.807

Regression Table of Model 3

By using the third model, we performed contrasts on the model predictions (Figure 1). To address our primary hypothesis, we contrast the mean personal total income of the group of white male with the mean personal total income of the group of women of color. This contrast let us to conclude that white males get \$19,696.76 more on average than women of color, and a difference of this magnitude is unlikely to occur due to chance alone since p-value is smaller than 0.001.

contrast	estimate	SE	df	t.ratio	p.value
Male White - Female POC	19696.76	838.4625	119564	23.49152	0

Contrast between Male White and Women of Color

For the secondary hypothesis, we built the fourth and fifth models based on what we found from the previous nested F-tests. Thus, both of the models have the interaction term (SEX * RACE) but we only allow the fifth model (Figure 2) to have the additional interaction on AGE (SEX * RACE * AGE).

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
119563	6.791664e+14	NA	NA	NA	NA
119560	6.788207e+14	3	345742827517	20.29844	0

Nested F-test between Model 4 and Model 5

The result of the nested F-test shows that the fifth model explains more variability in personal total income since the p-value is extremely small (3.812e-13) and is clearly less than 0.05. Thus, we have now shown that the interaction with AGE is necessary for our model in order to explain the data. Based on this result, we used the fifth model (or full model) to conduct the comparisons to compare the average personal total income of white males and women of color within three age group: early career, mid career, late career. The early career is the age 29.7 years, the mid career is the age of 44.4 years, and the late career is the age of 59.2 years.

contrast	AGE	estimate	SE	df	t.ratio	p.value
Male White - Female POC	29.70399	13080.20	1069.7657	119560	12.227162	0
Male White - Female POC	44.44979	13393.11	871.0449	119560	15.375918	0
Male White - Female POC	59.19559	13706.03	1454.1950	119560	9.425167	0

Contrast between White Male and Women of Color by Age (Early, Mid, and Late Career)

By looking at the table, we know that for early career, white males get 13,080.198 dollars more on average than women of color while white males get 13393.115 dollars more on average than women of color for mid-career. Furthermore, the contrast between white males and women of color in late career also shows that white males get 13706.031 dollars more on average than women of color. Please note that all of these differences of magnitudes are unlikely to occur due to chance alone since the p-values are smaller than 0.001.

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	49018.917	935.824	52.381	0.000	47184.718	50853.116
SEX: Female	-13198.028	1812.380	-7.282	0.000	-16750.264	-9645.793
RACE: POC	1643.687	1990.936	0.826	0.409	-2258.516	5545.890
AGE	1020.106	19.317	52.810	0.000	982.246	1057.966
SEX: Female:RACEPOC	-895.517	3564.569	-0.251	0.802	-7882.015	6090.981
SEX: Female:AGE	-238.869	38.318	-6.234	0.000	-313.971	-163.767
RACE: POC:AGE	144.639	44.840	3.226	0.001	56.753	232.524
SEX: Female:RACEPOC:AGE	73.010	82.082	0.889	0.374	-87.870	233.890

Regression Table of Model 5

However, we can observe that some of the coefficients are not significant for the full model. Based on the individual z-test, we found out that RACE: POC, SEX: Female:RACEPOC, and SEX: Female:RACEPOC:AGE are not significant to the model since they have large p-values over 0.05. This implies that the interaction of sex, race, and age is not necessary. Below is the interpretation of each term.

Intercept: The predicted total personal income of white male at age 0 is $49018.917 on average.

SEX:Female: The predicted total personal income of white female at age 0 is $13198.028 less than a predicted total personal income of white male.

RACE:POC: The predicted total personal income of POC male at age 0 is $1643.687 more than a predicted total personal income of white male.

AGE: For each additional year of age, the total personal income of white male increases by $1020.106, on average.

SEX:Female:RACE:POC: The predicted total personal income of Female POC at age 0 is $12449.86 (-13198.028 + 1643.687 - 895.517) less than a predicted total personal income of white male. This is the change in the change in intercept representing the effect of being both female and POC.

SEX:Female:AGE: For each additional year of age, the total personal income of white female decreases by $238.869 more than white male, on average.

RACE:POC:AGE: For each additional year of age, the total personal income of POC male increasese by $144.639 more than white male, on average.

SEX:Female:RACE:POC:AGE: For each additional year of age, the total personal income of POC female increases by $998.871 (1020.1 + -238.869 + 144.63 + 73.01) more than white male, on average. This is the change in the change in slope representing the effect of being both female and POC.

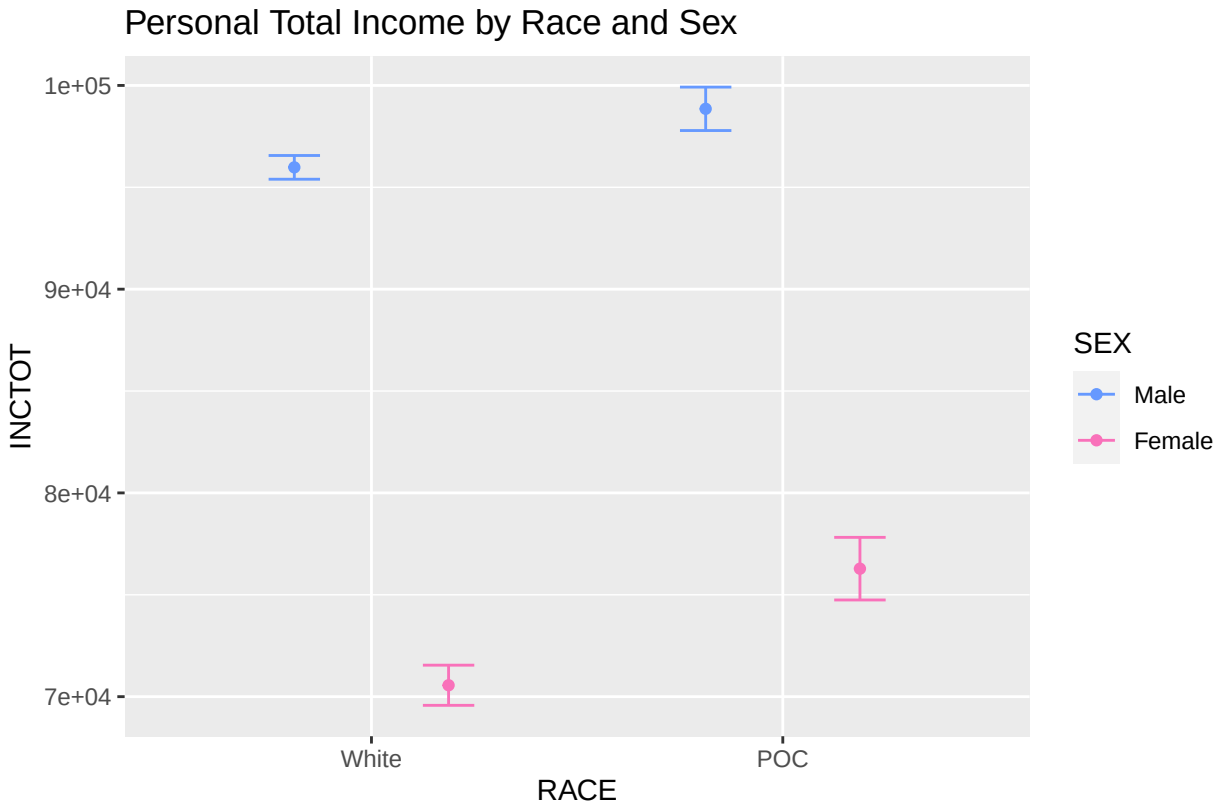


Figure 1: Contrast

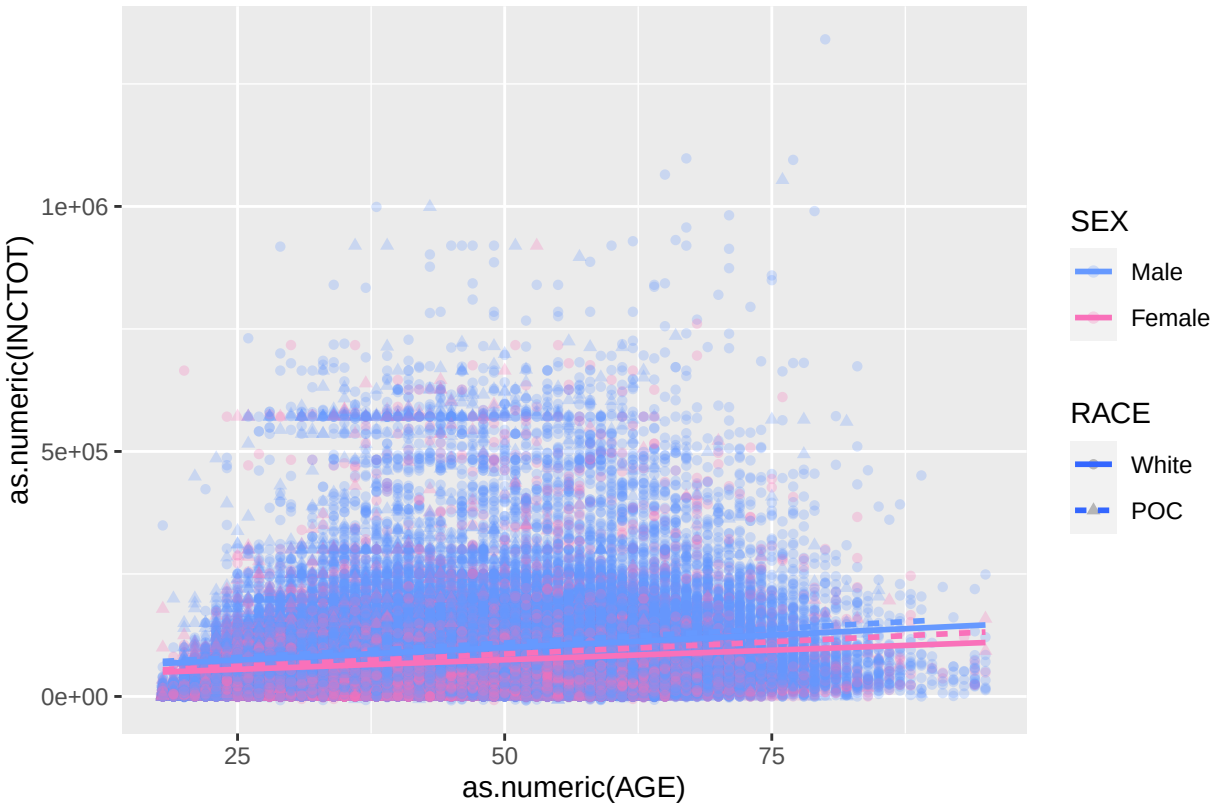


Figure 2: Model 5

Discussion

Our analysis successfully supports our primary hypothesis; among those who have STEM occupations, white male have higher total personal income compared to women of color, on average. We were also interested in seeing the effect of age on their personal total income and were able to observe that the difference between the mean total personal income of white male with SETM occupations and its of women of color with STEM occupations get slightly bigger as they get older. We, however, can not conclude that this is a significant difference since the interaction between sex, race, and age is not in necessary component of the interaction model (the fifth model).

Limitations

In this research, we only picked physical characteristics as our explanatory variables and did not consider other social status such as educational attainment and socio economic status. Moreover, we may have over simplified our sample by categorizing them into only two groups: White and POC. Additionaly, we filtered out all sample with missing data of total personal income. Our model may have lost valuable information from these points that may be missing because of reasons such as shame about income, flactuating income, reach to higher income, etc.

Since our models are categorical models, linearity is not applicable. Thus, we did not conduct linearity assumption test.

Conclusion

Our analysis shows that the combination of race and sex has significant effect on one's total personal income. However, age is not significantly associated with total personal income.

References

1. Funk, C., & Parker, K. (2019, December 31). Diversity in the stem workforce varies widely across jobs. Pew Research Center's Social & Demographic Trends Project. Retrieved December 11, 2021, from https://www.pewresearch.org/social-trends/2018/01/09/diversity-in-the-stem-workforce-varies-widely-across-jobs/.

2. Alan Zilberman and Lindsey Ice, "Why computer occupations are behind strong STEM employment growth in the 2019–29 decade," Beyond the Numbers: Employment & Unemployment, vol. 10, no. 1 (U.S. Bureau of Labor Statistics, January 2021), https://www.bls.gov/opub/btn/volume-10/why-computer-occupations-are-behind-strong-stem-employment-growth.htm

3. Xu, Y. (2015). Focusing on Women in STEM: A Longitudinal Examination of Gender-Based Earning Gap of College Graduates. The Journal of Higher Education 86(4), 489-523. doi:10.1353/jhe.2015.0020.

4. Searle Center for Advanced Learning & Teaching. Social Identities: Searle Center for Advancing Learning & Teaching. (n.d.). Retrieved December 11, 2021, from https://www.northwestern.edu/searle/initiatives/diversity-equity-inclusion/social-identities.html.

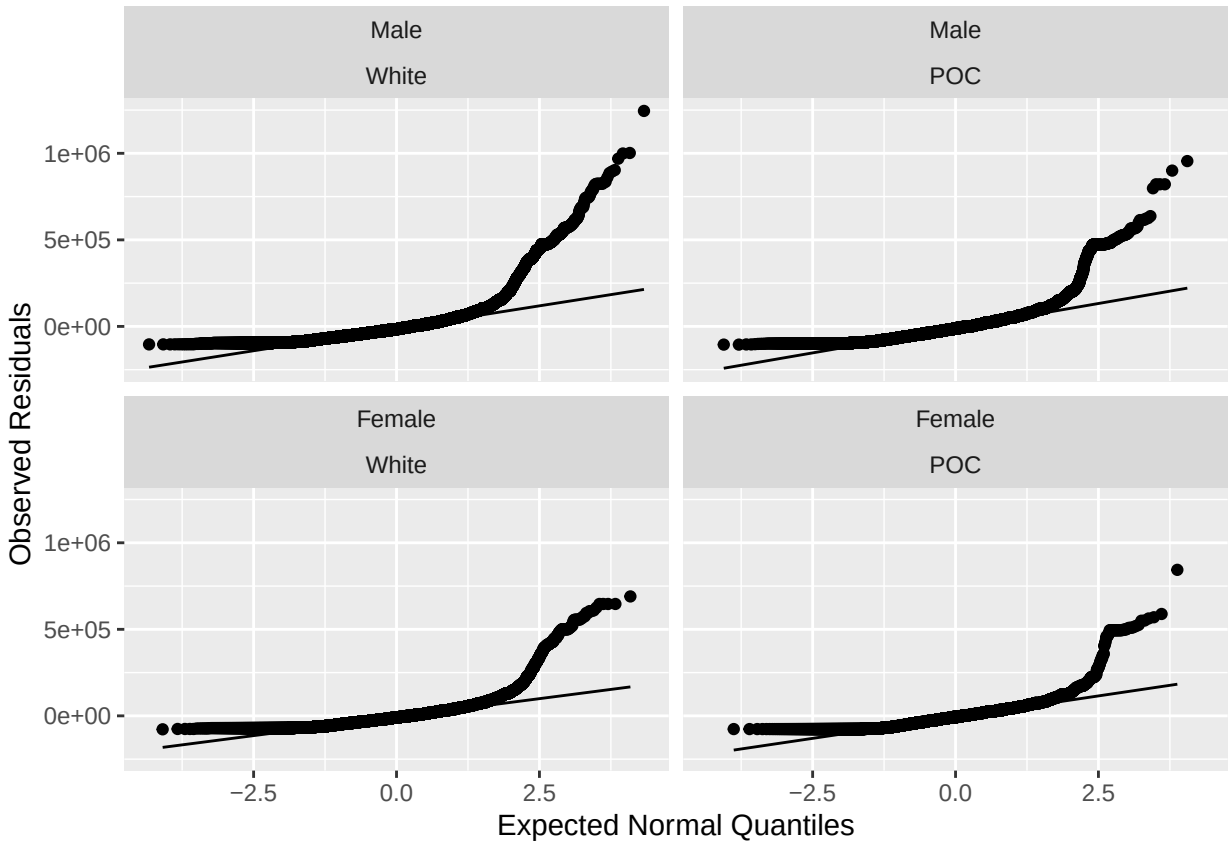
5. Suh, E. (2021, December 1). Age of majority by state as of 2021. Policygenius. Retrieved December 11, 2021, from https://www.policygenius.com/estate-planning/age-of-majority-by-state/.

Data Appendix

Checking L.I.N.E Assumptions

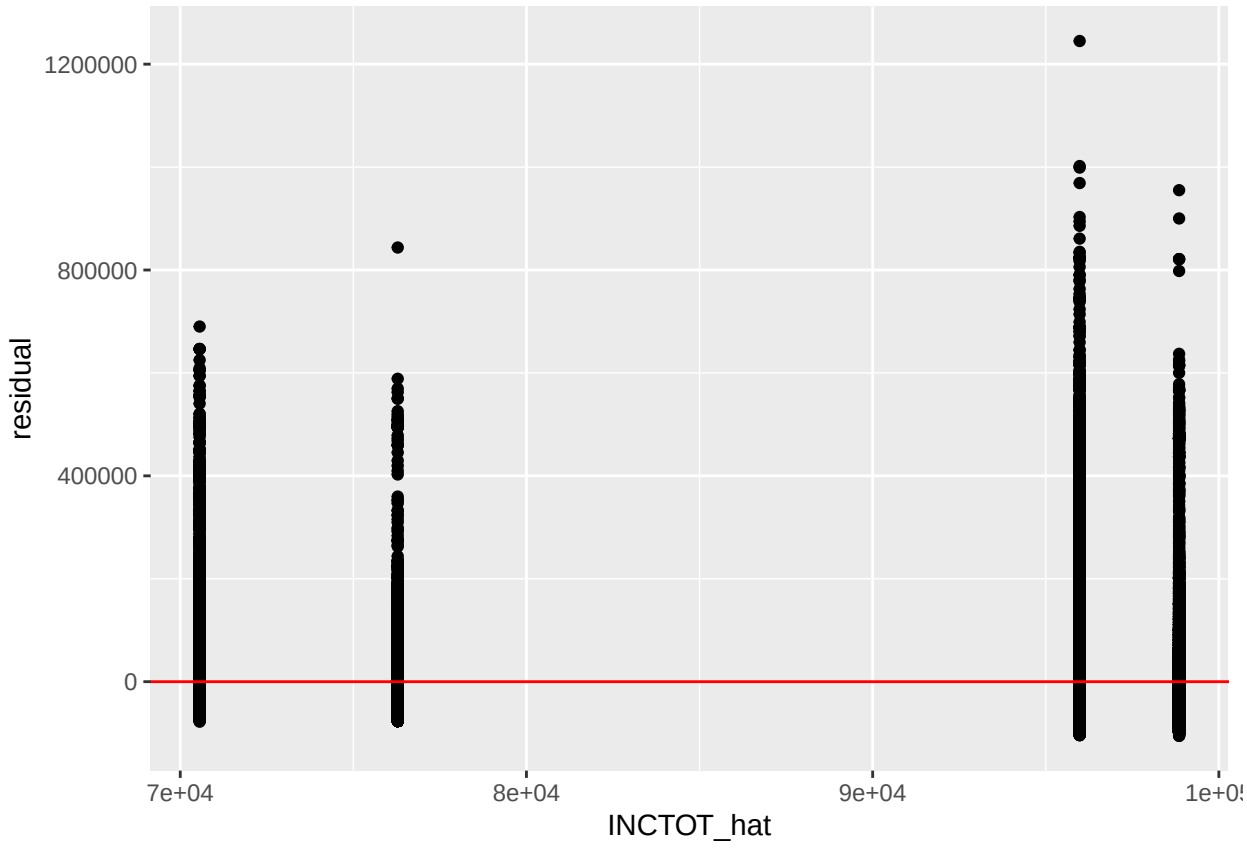
Model 3: `INCTOT ~ SEX * RACE` QQ Plot of the Residuals

Our model has four different groups: white male, white female, POC male, and POC female. Here, we show the QQ plot of the residuals of each group. Although the residual errors are clearly not normally distributed, this should not be the big concern since personal income should never be normally distributed.



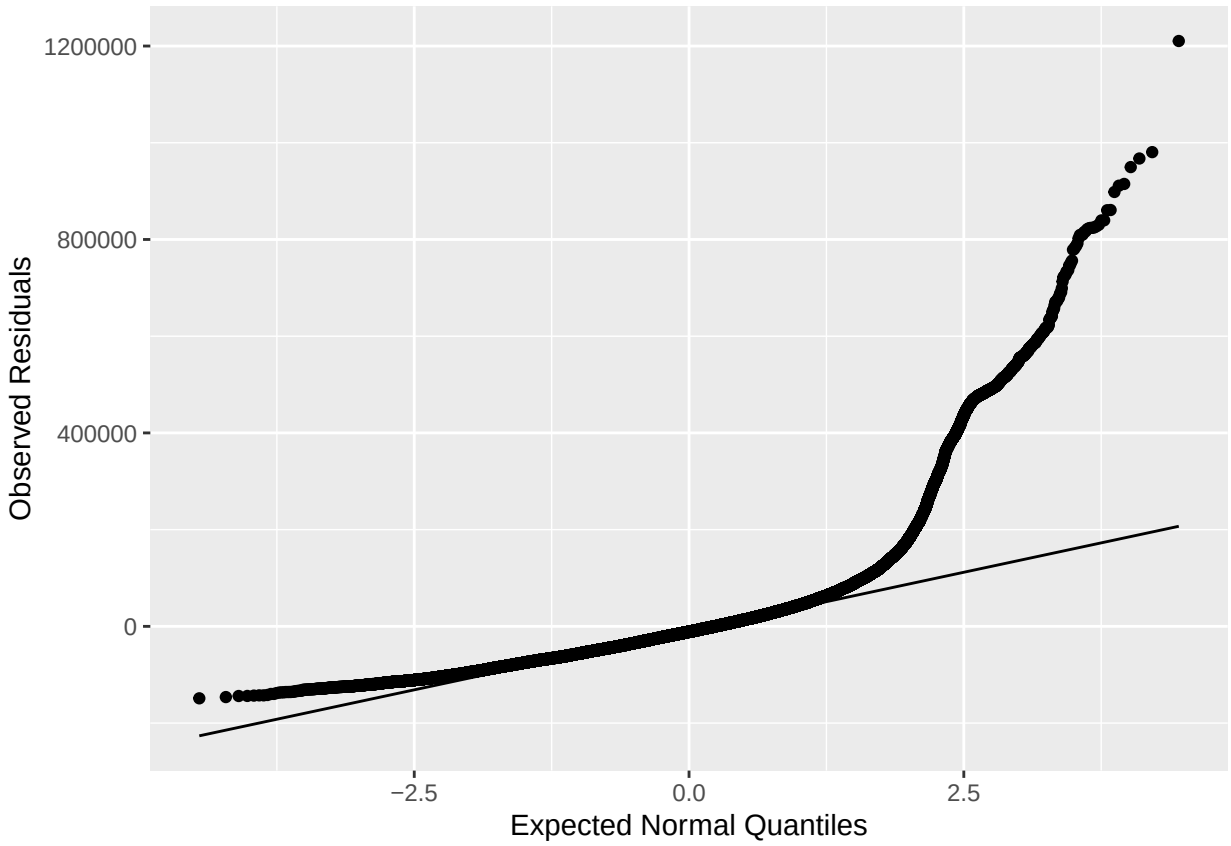
Equal Variance Assumption Test

We observed that the vertical clustering is pretty constant across the four groups. Thus, we conclude that equal variance assumption is met.



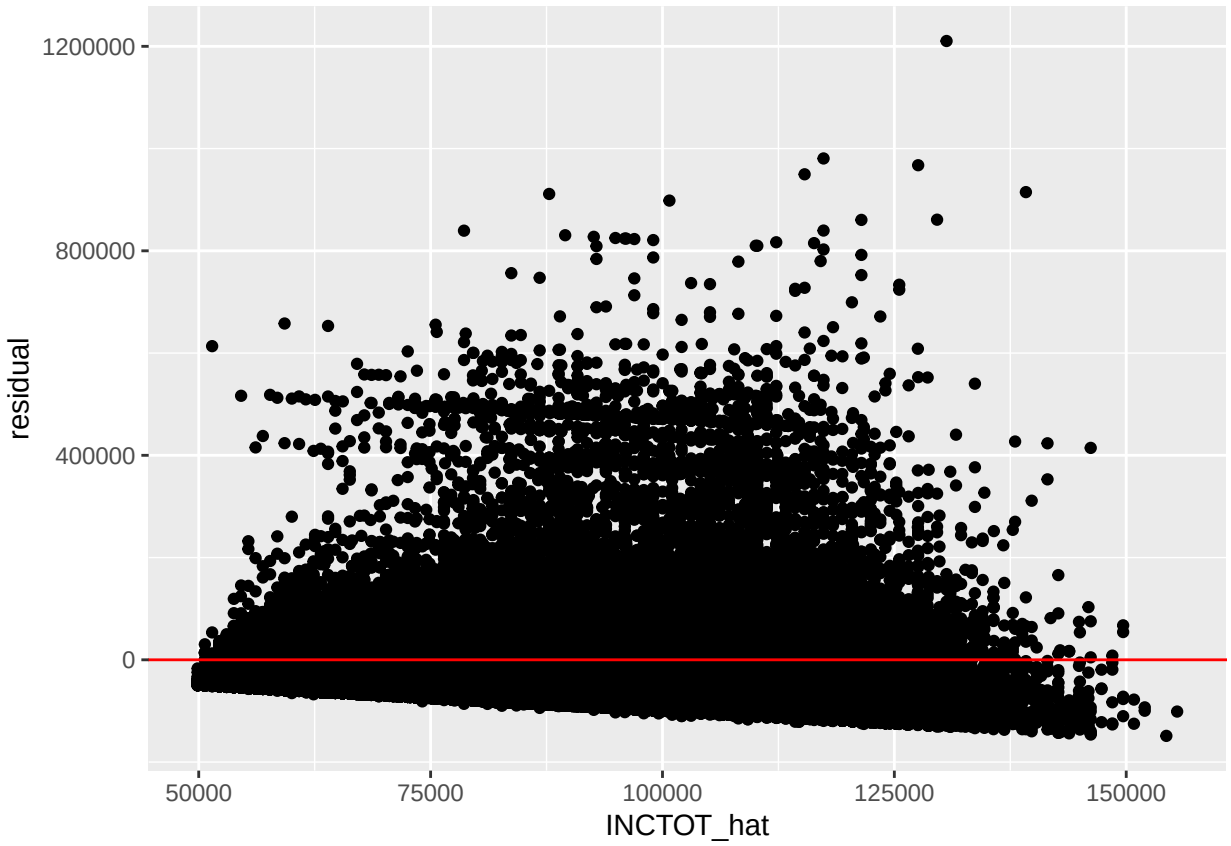
Model 5: `INCTOT ~ SEX * RACE * AGE` QQ Plot of the Residuals

For the same reason as we mentioned in the normality assumption test for model 3, this should not be the big concern.



Equal Variance Assumption Test

Equal variance is slightly violated but the overall trend of residuals match that of our residuals versus fitted values plot for the primary hypothesis. Therefore we will still include it in our secondary hypothesis as well.



Independence

As mentioned previously, our data has been randomly selected, each sampled only once, thereby satisfying the independence assumption.