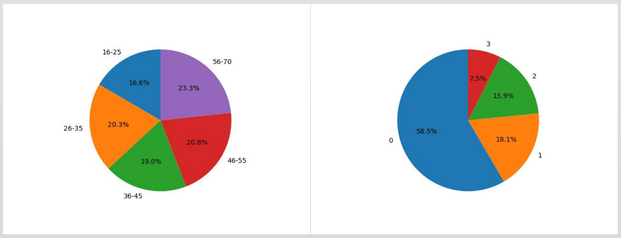
Do you think you are paid fairly?
An analysis of factors that affect annual salary

Hayden Truong and Haoxiang Zhang
Emory University Quantitative Methods Department

Introduction

Humans spend a third of our waking time working. Some spend even more. Yet it is unfortunate that the overall working experience is usually determined based on a couple of digits: one's annual salary. Looking to improve working conditions among adults, we study the factors that affect annual salaries. Which factors seem to affect income more than others, and to what degree? Which factors just happen to correlate with annual salary, and which factors actually drive income to change? Our most important research question, ultimately, is whether one's expertise dictates their salary more than his/her identity background such as race, age, and gender. We hope to gain insights into the labor market and wage payouts structure.



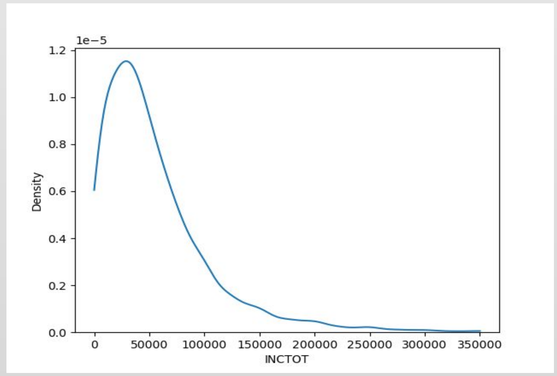
Distribution of age groups and number of children in the family among working-age people in 2021. Source: IPUMS

Literary Review

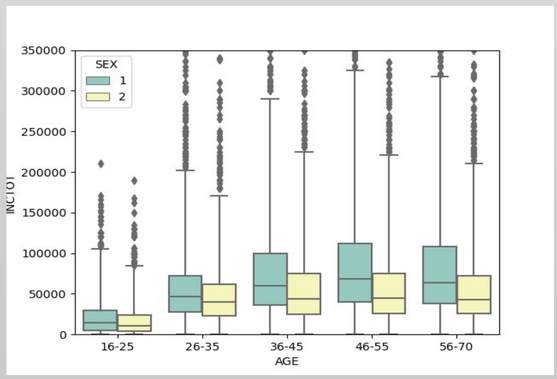
" In a regression comparing salary and non-monetary factors, only two out of 16 dummy variables are statistically significant. Both are positive work conditions, but Benefits (Positive) decreases salary and Long Hours (Positive) increases salary." (NYU, Mehta 2012)
"The main factors affecting the starting salary of graduates are the accumulation of human capital and social capital, but the segmentation of the labor market is also the main reason affecting the starting salary of graduates." (Wang 2022)

Methodology

The data investigation process utilized the data set from the IPUMS census bureau. By extracting 1000 random data , we were able to perform the following manipulations of data:
`Data Cleaning
·Data Visualization
·OLS Regressions
·Joint plots
Our research focus on the impact of different factors on income: particularly skillsets-related variables versus background-information variables, which most research has not covered fully.

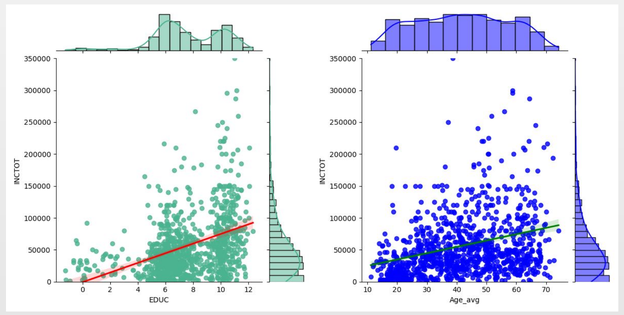


Income distribution among all respondents in 2021. (IPUMS)

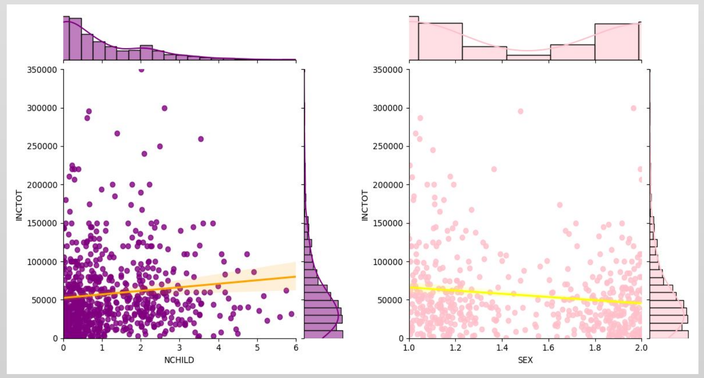


Income distribution groupby age and gender in 2021. (IPUMS)

Results



The above demonstrates regressions for education and age against total income. Education has a strong positive correlation with income and a decent slope; age also shows correlation. It can be indicated that higher education and age increase likelihood of higher personal income.



On the other hand, regression against more demographic variables like number of children and sex showed less correlation, which the low slope indicates that the factors have less impact on personal income. It can also be observed from the joint plot that the points are more scattered and random.

Conclusion

Regression results confirm our hypothesis that skillset-related variables such as education and industry impact income more than others. Regression slope and R-squared values for the above factors are higher than those of gender and race.
One weakness is that linear regression might not be the best fit model for our study as R-squared coefficients were not as high as could be. Although the trends in skillset variables were more apparent than other variables, the general correlation seems flat. Going forward, we hope to find a better model to run the data as well as suggesting policies that make the workplace more merit-based.

	Dependent variable:INCTOT
	(1)
SEX	-14241.569***
	(5120.300)
Age_avg	918.939***
	(198.087)
NCHILD	1886.411
	(2547.023)
MAR_STATUS	-2773.401**
	(1404.432)
CITIZEN	4704.846
	(3169.470)
RACWHT	18271.889***
	(6785.502)
Observations	1,000
R^2	0.068
Adjusted R^2	0.062
Residual Std. Error	80751.919 (df=993)
F Statistic	12.080*** (df=6; 993)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:INCTOT
	(1)
EDUC	12555.675***
	(1095.262)
IND	-2.184**
	(0.982)
SPEAKENG	4971.133
	(3583.212)
COGDIS	-19189.291
	(16029.046)
VISDIS	11777.729
	(15898.526)
Observations	1,000
R^2	0.123
Adjusted R^2	0.119
Residual Std. Error	78293.861 (df=994)
F Statistic	27.887*** (df=5; 994)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regression data for background variables (left) and skillset variables (right)

Acknowledgements

We would like to thank professor Dr. Palermo for his mentorship and for providing us the opportunity to conduct this study. We appreciate the Emory Library Services and Database for providing us the best resources. Finally, we thank our family and friends for being the best supporters.