ECONOMETRICS FINAL PROJECT

Matthew Evan Taruno

MAY 4, 2020

URID: 30677111

MW: 12:30 – 1:45PM

Worked on data cleaning with Alyssa Duan and James Sastrawan

CRITIQUE

INTRODUCTION

African immigrants increasingly make up a large part of the US economy, bringing billions of dollars from their spending and tax payments[6]. In terms of the overarching big picture, Ikpebe's research paper aims to investigate the effects of human capital and demographic variables on African immigrant earnings. In particular, to control for this effect, Ikpebe also used time related determinants, personal characteristics, and characteristics of immigrant country or origin as controls that make the regression model more accurate. Ikpebe uses these predictors because he he wants to quantify and capture the effects of cultural assimilation, show that some skills are not directly transferable from their home country, and show that immigrants eventually adapt and financially perform better when they acquire the US specific skillsets over time.

There are places where this paper falls short. For starters, based on a research paper by Bowles, he claims that years and level of schooling has surprisingly little effects on an individuals' earnings[1]. Rather, children of successful parents is a stronger predictor. However, this is not the focus of my paper. In my opinion, one of the bigger effect that is not accounted for by the model is a factor that directly effects immigrant earnings as well as native resident earnings. This factor is the *destination state* that immigrants end up choosing to settle at. Based on Table 3, it can be deduced that the country of origin effects – especially the ones relating to the colonization – are not too big and close to negligible and should be replaced with a more significant contributor to African immigrant earnings. Instead of country of origin effects, I believe that a stronger effect would be the effect of which state the immigrant immigrates to. This is my main point of critique, and I believe that this would be better to put as controls to add to the baseline model.

From the eyes of an immigrant, one of the biggest decisions they will have to make is which state to immigrate towards. An immigrant might choose to immigrate to the US may do so because of less financial opportunities in their home country, to gain advanced knowledge to bring back to their country of origin, or even start a successful small business in the streets of Chinatown[1]. This decision has a definite effect on the level of income they will ultimately make due to a variety of factors such as cost of living, transportation, minimum wage laws, and proportion of tax contribution to name a few[1]. In the regression model, all of these factors that effect level of income could potentially be used as proxy variables to account for the bias caused from omitting the effects of destination state on real wage. Ikpebe's analysis does not modify the model at all to account for these important effects. I will further discuss the particular bias effects on the evidence section of this paper.

DATA AND REPLICATION

Here are my two replication tables:

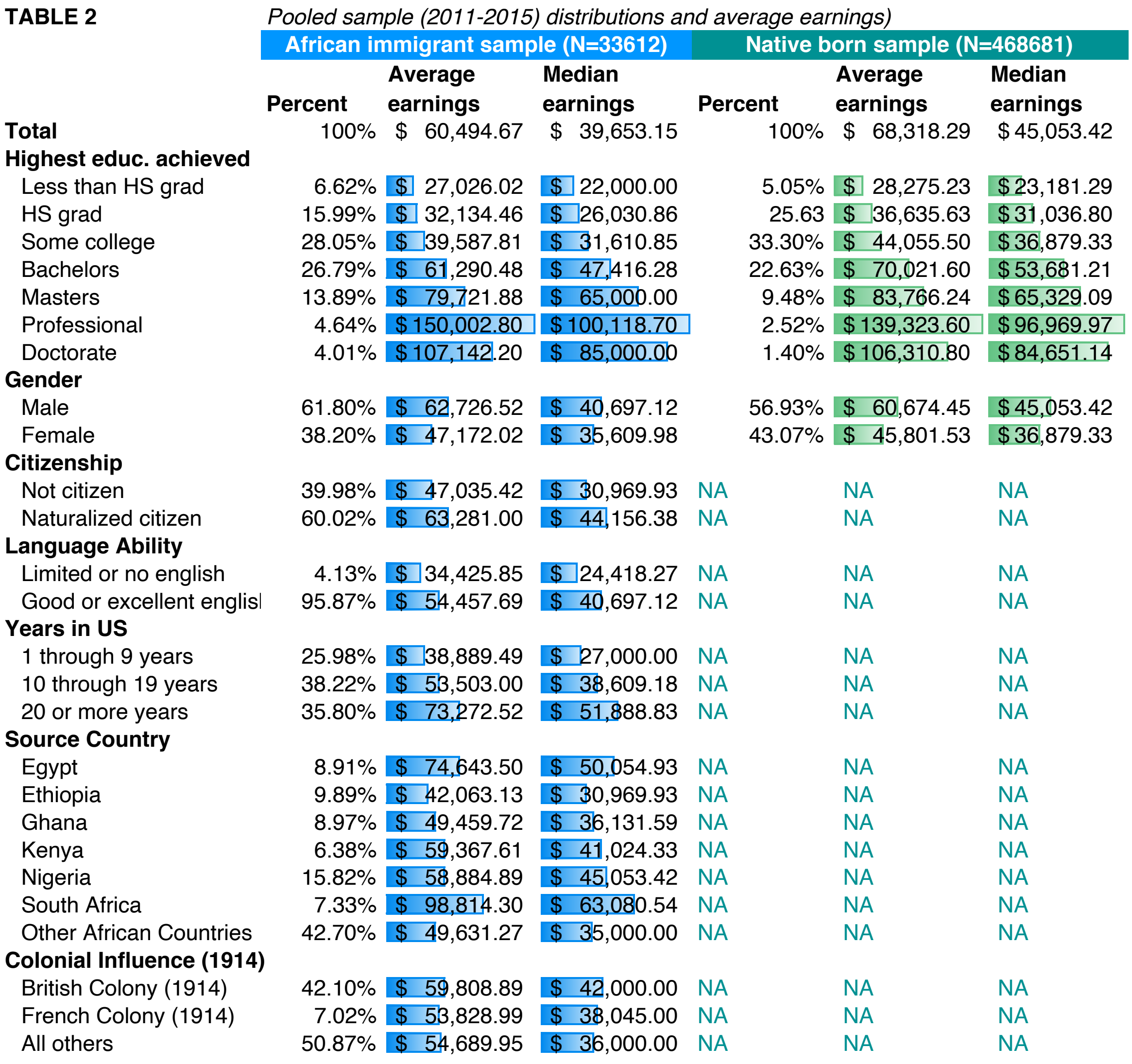


TABLE 2	*Pooled sample (2011-2015) distributions and average earnings)*					
	African immigrant sample (N=33612)			Native born sample (N=468681)		
		Average	Median		Average	Median
	Percent	earnings	earnings	Percent	earnings	earnings
Total	100%	$ 60,494.67	$ 39,653.15	100%	$ 68,318.29	$ 45,053.42
Highest educ. achieved						
Less than HS grad	6.62%	$ 27,026.02	$ 22,000.00	5.05%	$ 28,275.23	$ 23,181.29
HS grad	15.99%	$ 32,134.46	$ 26,030.86	25.63	$ 36,635.63	$ 31,036.80
Some college	28.05%	$ 39,587.81	$ 31,610.85	33.30%	$ 44,055.50	$ 36,879.33
Bachelors	26.79%	$ 61,290.48	$ 47,416.28	22.63%	$ 70,021.60	$ 53,681.21
Masters	13.89%	$ 79,721.88	$ 65,000.00	9.48%	$ 83,766.24	$ 65,329.09
Professional	4.64%	$ 150,002.80	$ 100,118.70	2.52%	$ 139,323.60	$ 96,969.97
Doctorate	4.01%	$ 107,142.20	$ 85,000.00	1.40%	$ 106,310.80	$ 84,651.14
Gender						
Male	61.80%	$ 62,726.52	$ 40,697.12	56.93%	$ 60,674.45	$ 45,053.42
Female	38.20%	$ 47,172.02	$ 35,609.98	43.07%	$ 45,801.53	$ 36,879.33
Citizenship						
Not citizen	39.98%	$ 47,035.42	$ 30,969.93	NA	NA	NA
Naturalized citizen	60.02%	$ 63,281.00	$ 44,156.38	NA	NA	NA
Language Ability						
Limited or no english	4.13%	$ 34,425.85	$ 24,418.27	NA	NA	NA
Good or excellent englisl	95.87%	$ 54,457.69	$ 40,697.12	NA	NA	NA
Years in US						
1 through 9 years	25.98%	$ 38,889.49	$ 27,000.00	NA	NA	NA
10 through 19 years	38.22%	$ 53,503.00	$ 38,609.18	NA	NA	NA
20 or more years	35.80%	$ 73,272.52	$ 51,888.83	NA	NA	NA
Source Country						
Egypt	8.91%	$ 74,643.50	$ 50,054.93	NA	NA	NA
Ethiopia	9.89%	$ 42,063.13	$ 30,969.93	NA	NA	NA
Ghana	8.97%	$ 49,459.72	$ 36,131.59	NA	NA	NA
Kenya	6.38%	$ 59,367.61	$ 41,024.33	NA	NA	NA
Nigeria	15.82%	$ 58,884.89	$ 45,053.42	NA	NA	NA
South Africa	7.33%	$ 98,814.30	$ 63,080.54	NA	NA	NA
Other African Countries	42.70%	$ 49,631.27	$ 35,000.00	NA	NA	NA
Colonial Influence (1914)						
British Colony (1914)	42.10%	$ 59,808.89	$ 42,000.00	NA	NA	NA
French Colony (1914)	7.02%	$ 53,828.99	$ 38,045.00	NA	NA	NA
All others	50.87%	$ 54,689.95	$ 36,000.00	NA	NA	NA

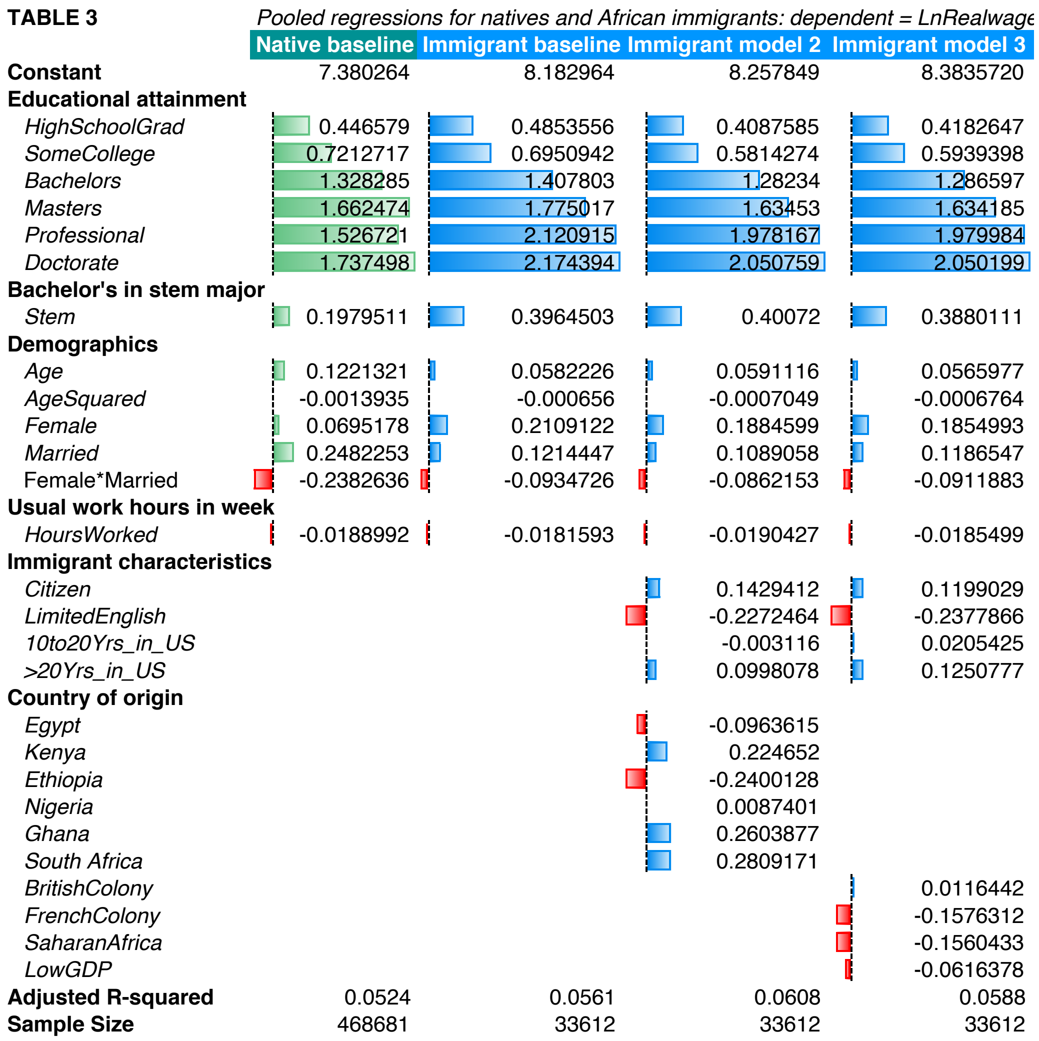


TABLE 3 *Pooled regressions for natives and African immigrants: dependent = LnRealwage*

	Native baseline	Immigrant baseline	Immigrant model 2	Immigrant model 3
Constant	7.380264	8.182964	8.257849	8.3835720
Educational attainment				
HighSchoolGrad	0.446579	0.4853556	0.4087585	0.4182647
SomeCollege	0.7212717	0.6950942	0.5814274	0.5939398
Bachelors	1.328285	1.407803	1.28234	1.286597
Masters	1.662474	1.775017	1.63453	1.634185
Professional	1.526721	2.120915	1.978167	1.979984
Doctorate	1.737498	2.174394	2.050759	2.050199
Bachelor's in stem major				
Stem	0.1979511	0.3964503	0.40072	0.3880111
Demographics				
Age	0.1221321	0.0582226	0.0591116	0.0565977
AgeSquared	-0.0013935	-0.000656	-0.0007049	-0.0006764
Female	0.0695178	0.2109122	0.1884599	0.1854993
Married	0.2482253	0.1214447	0.1089058	0.1186547
Female*Married	-0.2382636	-0.0934726	-0.0862153	-0.0911883
Usual work hours in week				
HoursWorked	-0.0188992	-0.0181593	-0.0190427	-0.0185499
Immigrant characteristics				
Citizen			0.1429412	0.1199029
LimitedEnglish			-0.2272464	-0.2377866
10to20Yrs_in_US			-0.003116	0.0205425
>20Yrs_in_US			0.0998078	0.1250777
Country of origin				
Egypt			-0.0963615	
Kenya			0.224652	
Ethiopia			-0.2400128	
Nigeria			0.0087401	
Ghana			0.2603877	
South Africa			0.2809171	
BritishColony				0.0116442
FrenchColony				-0.1576312
SaharanAfrica				-0.1560433
LowGDP				-0.0616378
Adjusted R-squared	0.0524	0.0561	0.0608	0.0588
Sample Size	468681	33612	33612	33612

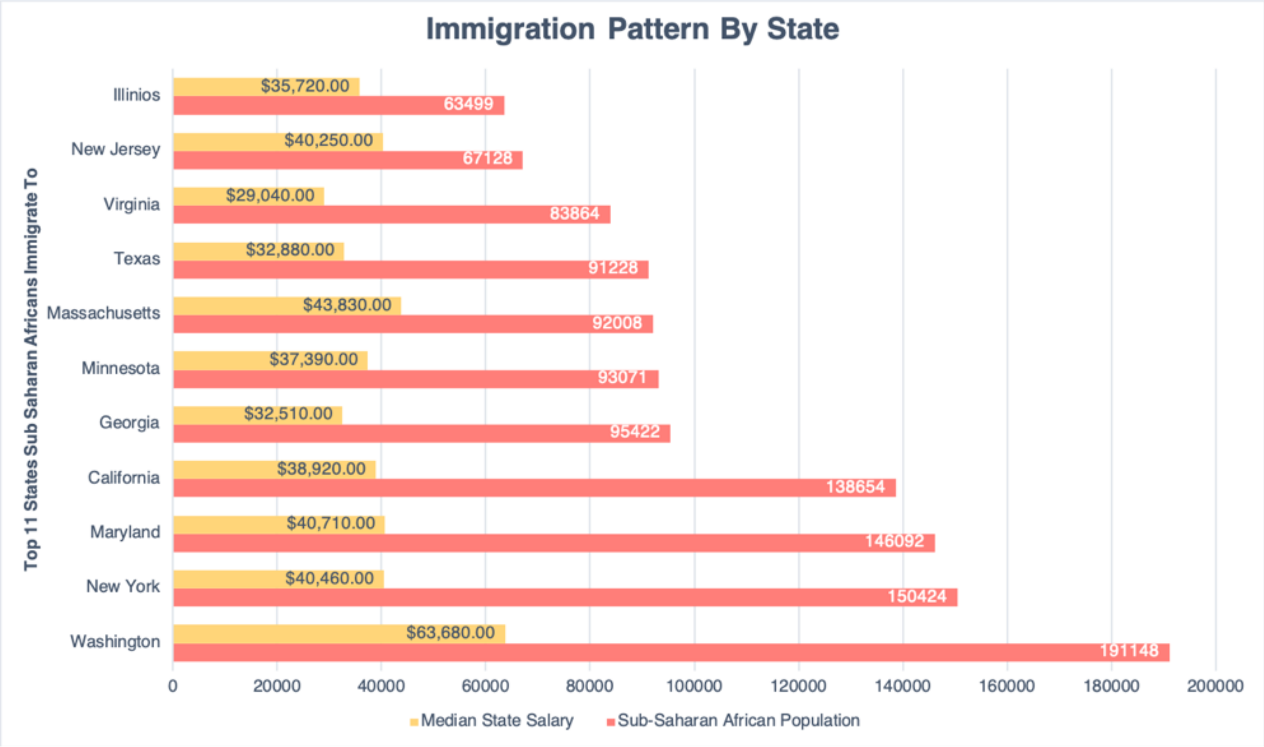
In order to replicate the tables, I went through an extensive process of cleaning the data – this entire process can be seen from the do file that I attached to this final project. First, I dropped all the observations that are not in the range from 2011-2015. I also dropped all observations that had less than or equal to 35 hours of work. Then I adjusted all the wages for inflation by taking the annual CPI averages and converting each year to the 2015 value by calculating the ratio and multiplying this ratio to all the observations that match the respective year. For example, the ratio for 2011 was 1.053695. I also generated a new column that is the natural log of these real wages. Next, since there are more categories in the data than are necessary and they are all in detail, I made sure to bin these categories in accordance to the data we have in table 2 and 3 in Ikpebe's paper. In order to do this, I realized that even though the categories in Stata display Strings, each category corresponded to a number. Therefore, for each of the categories of interest, I used a series of Boolean math to scrape out the categories of interest and convert them into my own numerical categories. For example, for the years of schooling, 1000 corresponded to less than HS graduates, 2000 to HS graduates, and so on up until 7000 which represented a doctorate. A challenge I encountered was that sometimes, my numerical categories corresponded to the predetermined numerical category, however I overcame this by restarting Stata and backtracking my Stata code to where my problem conversion lied and used new numerical categories. In order to make sure I isolated the immigrants, I used an if statement to filter out the data on Africans, which had a code of 600.

For table 3, I continued with the category replacements I applied for table 2 onto new variables. After this, I dummy encoded all the respective variables for k-1 categories. I encountered firsthand that if I did not use k-1 categories, when I ran the regression I would always have one predictor have a 0 coefficient because of multicollinearity. In order to get the cumulative percentages I tabulated the variable, and for getting the mean and median I used a detailed expansion of Stata's summary feature and took the 50th percentile as the median. The rest of my transformations are straightforward and are extensively captured in my do file.

EVIDENCE

As alluded to in the introduction, this section will expand on why destination state is an important omitted effect to consider.

First things first, we can check that destination state is an omitted variable by checking if it is a confounder. We know this true because being in a higher salary state could lead to higher wages (vice versa), and some states are more suitable for older people as well as more suitable for married couples so human capital related factors are effected by location as well. Additionally, no other regressor in the model accounts for this effect like how education level partially accounts for the level of intellectual ability, so we can assume that state destination is not redundant.



This visualization gives a good idea of where sub-saharan Africans immigrate to and how much money (in terms of median salaries) could be made in the states these immigrants immigrate to[3,6]. All together, the population numbers above make 75.9% of the total sub-saharan African immigrant population, so it is safe to say that this is a good representation of the sub-saharan African immigrant population in the United States. In order to roughly get the bias you would receive (whether or not it is an upward or downward bias) from going to a particular state, it might be useful to compare the median wage you could obtain from a particular state to the expected value of how much immigrants are making in general in the US. If the median wage in the state is greater than the expected general value for income wage from the research paper, than we can generally assume that there will be an upward bias.

In order to better obtain the quantitative effect of destination state on the regression model, I tried to look for data that could be inputted into the regression model in Stata. However, since I could not find corresponding data to where the immigrant has immigrated towards, so I decided to create my own data through a simulation. Based on the table below (or the visualization above from that matter), I made a probability distribution of how likely an immigrant would be in a particular state by simply taking the population of immigrants in a particular state divided by the sum of the entire African immigrant population. The probabilities of a particular observation for my sampling to have a particular median state salary are shown below in the probability column:

States	Sub Saharan African Population	Median State Salary	Probability
Washington	191148	$63,680	15.76%
New York	150424	$40,460	12.41%
Maryland	146092	$40,710	12.05%
California	138654	$38,920	11.44%
Georgia	95422	$32,510	7.87%
Minnesota	93071	$37,390	7.68%
Massachusetts	92008	$43,830	7.59%
Texas	91228	$32,880	7.52%
Virginia	83864	$29,040	6.92%
New Jersey	67128	$40,250	5.54%
Illinois	63499	$35,720	5.24%

I sampled from this probability distribution and assigned the median salary from the state selected by the distribution at every observation until the entire column called simul_wage was created. Then I log transformed this data to make it comparable to the lnrealwage response variable that is used by the research paper and successfully created a variable called ln_simul_wage using Python. Then I converted the csv file back into a Stata dta file, then I used the merge by observations edit mode in Stata to add the simul_wage column to add back to Stata to obtain this dataset. With this variable, I used it as a predictor variable in my regression model to account for the effects of omitting destination state effects, which is my critique of the paper. While this methodology is not perfect, as I will address to below, it does capture some quantitative sense of the magnitude and direction of the bias caused by my omitting these effects. The results of the regression is shown below in Table 4 in purple. I recalculated and ran the regression with the inclusion of destination state effect that corresponded to the immigrant baseline model and the expanded immigrant model 2 represented in blue. The idea is to have, on an aggregate population level, the expected distribution effect of median wages in the data and to account for destination state effects in the data.

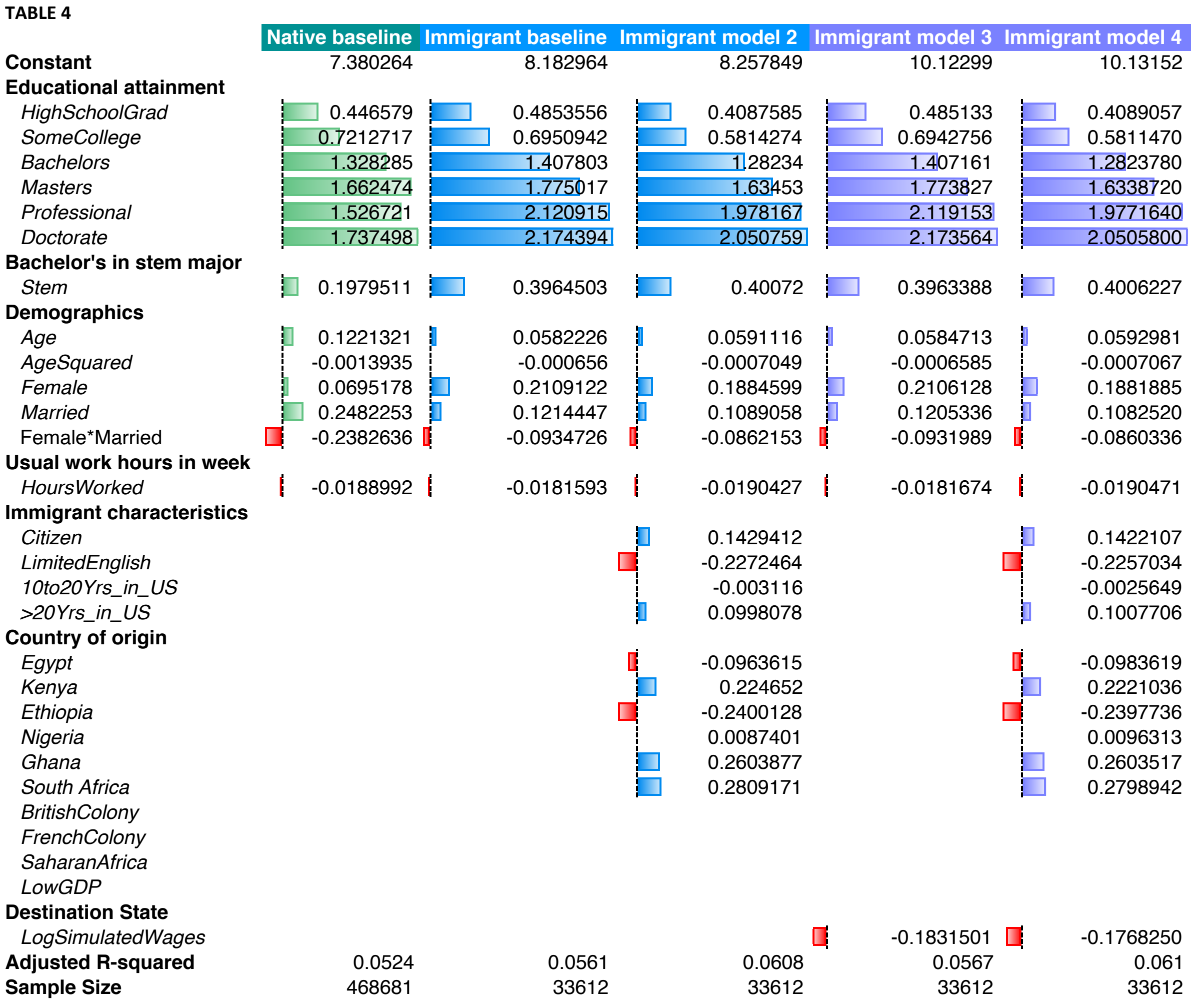


TABLE 4

	Native baseline	Immigrant baseline	Immigrant model 2	Immigrant model 3	Immigrant model 4
Constant	7.380264	8.182964	8.257849	10.12299	10.13152
Educational attainment					
HighSchoolGrad	0.446579	0.4853556	0.4087585	0.485133	0.4089057
SomeCollege	0.7212717	0.6950942	0.5814274	0.6942756	0.5811470
Bachelors	1.328285	1.407803	1.28234	1.407161	1.2823780
Masters	1.662474	1.775017	1.63453	1.773827	1.6338720
Professional	1.526721	2.120915	1.978167	2.119153	1.9771640
Doctorate	1.737498	2.174394	2.050759	2.173564	2.0505800
Bachelor's in stem major					
Stem	0.1979511	0.3964503	0.40072	0.3963388	0.4006227
Demographics					
Age	0.1221321	0.0582226	0.0591116	0.0584713	0.0592981
AgeSquared	-0.0013935	-0.000656	-0.0007049	-0.0006585	-0.0007067
Female	0.0695178	0.2109122	0.1884599	0.2106128	0.1881885
Married	0.2482253	0.1214447	0.1089058	0.1205336	0.1082520
Female*Married	-0.2382636	-0.0934726	-0.0862153	-0.0931989	-0.0860336
Usual work hours in week					
HoursWorked	-0.0188992	-0.0181593	-0.0190427	-0.0181674	-0.0190471
Immigrant characteristics					
Citizen			0.1429412		0.1422107
LimitedEnglish			-0.2272464		-0.2257034
10to20Yrs_in_US			-0.003116		-0.0025649
>20Yrs_in_US			0.0998078		0.1007706
Country of origin					
Egypt			-0.0963615		-0.0983619
Kenya			0.224652		0.2221036
Ethiopia			-0.2400128		-0.2397736
Nigeria			0.0087401		0.0096313
Ghana			0.2603877		0.2603517
South Africa			0.2809171		0.2798942
BritishColony					
FrenchColony					
SaharanAfrica					
LowGDP					
Destination State					
LogSimulatedWages				-0.1831501	-0.1768250
Adjusted R-squared	0.0524	0.0561	0.0608	0.0567	0.061
Sample Size	468681	33612	33612	33612	33612

Several things come into light: the coefficient of log simulated wages is roughly a slight negative effect, of -0.18. For a unit change in the wages, we expect to see the predicted wage to decrease. Or more simply put, the overall effect of having destination state effects accounted for in the model is that we have a slight negative effect on the expected earnings of an immigrant.

The bias should have two components. Beta w, which concerns the relationship between real wage and destination state. And theta hat which concerns the relationship between human capital factors and destination state. Beta w could be positive if in general, more immigrants in the sample end up in states where they can earn more. It would be negative if in general, more immigrants can be found in states where they earn less than the expected earnings. Based on the table containing the probabilities, we get the median of all median state salaries are $38,920. We see that from the dataset, most immigrants earn more than this median, so I am going to deduce that the overall effect of beta w is positive. As for theta hat, different states have different human capital demographics, so the overall effect can be positive in the case where in general, immigrants move to states with a higher older population. Or if they move to a state where they are more likely to be female and married. Since we expect beta w to be positive, we can deduce that theta hat is negative because the coefficient of log simulated wages is negative.

However, there is also a likely scenario where theta hat could be positive. One thing to note in relation to the given paper is that Ikpebe mentions how there is an overrepresentation of African immigrants who go to the STEM field. This will be important to note, because STEM majors tend to earn more than non-STEM majors, and not by a small amount. Based on the US Bureau of Labor Statistics, STEM occupations make a median annual wage in 2019 of $86,980 compared to non-STEM occupations that made $38,160[4]. This means that the wages of the immigrant sample is inflated on the Ikpebe's research paper, even though the top occupations for African immigrants where they are more likely to take a job over a US native are nursing, housekeeping, and taxi drivers among the lower paid occupations. Among the higher paid occupations, Africans lean into being registered nurses, postsecondary teachers, and being surgeons[6]. The effect this has on my simulation is that there is an underestimate for the wages, so in actuality, the coefficient of log simulated wages is supposed to be positive.

Nevertheless, regardless of the sign of the bias, we still arrive at the conclusion that omitting destination state effects causes a significant change to the model. We can take a closer look at this effect by investigating immigrant models 3 and 4 in purple in Table 4. The absolute magnitude of 0.18 is greater than the effect of spending 20+ years in the US which has a magnitude of 0.1. It is also greater than the demographic factors age and marital status.

CONCLUSION

In conclusion, Ikpebe has done a good job capturing some of the main effects of immigrant wage earnings. However, the effect of the which state the immigrant decides to immigrate to is not something that Ikpebe takes into account, which introduces omitted variable bias into the regression model. I argue that we should look at this problem statement of trying to find the causal effect of human related factors on earnings from the eyes of the immigrant. Imagine you are trying to immigrate to the US from an African country, one of your major considerations will definitely be to carefully choose where you want to live because different states have different business districts, opportunities, educational climates, or simply to find a safe and affordable living standards. The Pew Research Center finds that another reason for this major influx of African immigrants since the 20th century is Refugee Act of 1980, which exposed the conflict of areas such as Somalia and Ethiopia[8]. This drive to find a safer state where you can get a sustainable living is one of the reasons why destination state plays a big part on your earnings. Overall, I demonstrated that this effect is slightly negative from the regression that I ran in an attempt to simulate and quantify these state effects.

For future research, I can make my simulation utilize median wages of STEM graduates because Ikpebe oversampled immigrants who are in the field of STEM. Moreover, even better, I could get more updated data that accounts for all Ikpebe's and investigate which state each of the immigrants go to. This will better isolate destination state effects has on the regression and help us determine the particular nature of the overall bias. However, as discussed before, regardless of the nature of the bias, we can conclude that there is evidence that destination state plays a major enough role to account it into the regression model, even over what was thought to be important factors such as time spent in the US that Ikpebe used to capture cultural assimilation effects.

BIBLIOGRAPHY

1. “African Immigration to the United States.” Wikipedia. Wikimedia Foundation, May 1, 2020. https://en.wikipedia.org/wiki/African_immigration_to_the_United_States#Demographic.
2. “Best 5 States to Immigrate to in the US.” USDV Experts. Accessed May 4, 2020. https://usdvexperts.com/language/en/best-5-states-to-immigrate-to-in-the-us/.
3. Bowles, Samuel. “The Determinants of Earnings: A Behavioral Approach.” UMass, January 26, 2001. https://www.umass.edu/preferen/gintis/jelpap.pdf.
4. “Employment in STEM Occupations.” U.S. Bureau of Labor Statistics, April 15, 2020. https://www.bls.gov/emp/tables/stem-employment.htm.
5. “Historical Consumer Price Index for All Urban Customers.” Accessed May 4, 2020. https://www.bls.gov/cpi/tables/historical-cpi-u-201709.pdf.
6. “How Sub-Saharan Africans Contribute to the US Economy.” New American Economy, January 2018. https://research.newamericaneconomy.org/wp-content/uploads/sites/2/2018/01/NAE_African_V6.pdf.
7. “Salary By State: Where Can You Really Earn the Most?” Rasmussen. Accessed May 4, 2020. https://www.rasmussen.edu/student-experience/college-life/salary-by-state/.
8. Anderson, Monica. “African Immigrant Population in U.S. Steadily Climbs.” Pew Research Center, February 14, 2017. https://www.pewresearch.org/fact-tank/2017/02/14/african-immigrant-population-in-u-s-steadily-climbs/.