Do States With High Income Also Have More Immigrants?

Niall Gamble, Adrian Campos Suarez, William Yao

ECON 3161: Econometric Analysis

Dr. Shatakshee Dongde

Fall 2019

Abstract

Data was collected from the 50 states and the Capital of the United States of America in 2017 for evaluation. The main analysis investigates whether a state's median household income is positively correlated with the state's immigration. While a simple linear model showed a positive correlation. Subsequent models did not show the same relationship between a state's immigration and a state's income, and instead showed state immigration to be positively related to the presence of Fortune 500 Companies and Top 100 Universities. Future investigation should go into explaining the negative relationship.

1. Introduction

In various economic models, economists put forth the assumption that individuals act in their own self-interest. To some extent, the main assumption underlying capitalistic thought is that: individuals acting in their own self-interests will lead to greater benefits for society as a whole. Under these kinds of assumptions, one would expect individuals to seek the most return on their investments in all aspects of life by determining how much they are willing to pay for a given good or service. Or in this situation, their choice of which job to take based on the expected income and other associated benefits for the time which they give up. However, these assumptions which are made in the confines of models may not always prove true in the real world. Because of this discrepancy, economists seek to put their models and their assumptions to test in the real world by comparing their theoretical models to real world statistics and surveys. Naturally, this study seeks to do the same regarding the decisions that individuals make when looking for a place for them to live.

This paper will analyze the effects of a state's median household income on the levels of immigration. As put forth earlier, individuals seek out the best opportunities for themselves, which in this situation would be determining where to live. Individuals take into account the potential of finding a job with a high marginal benefit for them, and this paper assumes that a state with high median household income would be seen by individuals as having a higher marginal benefit for them then a state with lower median household income. This relationship which is being implied would be reflected statistically in records regarding the amount of immigration into a state. The result which this paper expects is for states with higher median household income will have greater levels of immigration in a given year since they would be more attractive to move when an individual is making a decision on where to live in the future when compared to states with lower median household incomes which are less desirable when an individual is making a decision on where to live in the future. The core question which this paper seeks to answer is: Do states with higher levels of immigration appear more attractive to states with less immigration because of their higher median household income, which individuals interpret as higher rate return of money for their time?

As stated earlier, this paper is analyzing this relationship by testing several economic assumptions which have been put forth and accepted in theoretical models. Specifically, the paper is testing the assumption which puts forth that individuals behave rationally and make decisions based on the marginal benefit and opportunity costs of the choices they have available to them. This paper is also testing the assumptions in many supply-side models which state that labor, which consists of individuals, is variable and can quickly adapt to changes in demand for more or less labor in various locations and industries.

2. Literature Review

A key part of studies not just in economics but in all fields involves on building upon the work done by previous studies. To this end, this paper will now address the previous works regarding the relationships between immigration, state median income, and the regional discrepancies thereof.

Cebula and Alexander (2006) performed a similar analysis of this issue similar to the one this study is doing. The model which was being tested took into account the median family income from the 50 states along with a few other variables to serve as controls in order to predict the net-migration rate in those states. The key difference between this paper and the Cebula and Alexander (2006) paper is the type of analysis. While this paper seeks to do a cross-sectional data analysis the Cebula and Alexander (2006) paper performed a time series analysis from 2000 to 2004. A key point brought up by the Cebula Alexander (2006) paper is the importance of including non-economic factors when taking into consideration the possible factors that could have an effect on the net-immigration in a state. As the paper puts it "omission of non-economic factors from an empirical migration analysis constitutes an omitted-variable problem that generally compromises the integrity of that analysis" (Cebula and Alexander). For that reason, this papers research will also consider non-economic "quality of life" factors when creating its model for predicting net-immigration. Cebula and Alexander (2006) did successfully manage to show the effect of climate on the immigration rate showing the normal daily maximum temperature in January did have a statistically significant effect on the net-immigration even at the 1.0 percent level. The paper also showed that hazardous waste and toxic chemical released in a state to have a statistically significant effect on the net-immigration; however, this was only at the five percent level. Furthermore, Cebula and Alexander (2006) failed to show effects for geography on the net-immigration of a state since both the dummy variable and numerical variable that the paper used in their models failed to show a statistical significance at even the

10.0 percent level. Furthermore, Cebula and Alexander (2006) also showed the significance of other economic variables including: state income tax, cost of living, percent employment growth, along with the median family income at the 5.0 percent level for percent employment growth and 1.0 percent level for all the other variables. The findings from Cebula and Alexander (2006) help drive the direction of the variables that this paper will take into consideration and which variables this paper will avoid.

In addition to Cebula and Alexander's (2006) international standpoint, is Kazakis and Faggian's (2017) U.S. graduates' standpoint. Kazakis and Faggian (2017) explores the relationship between inter-state migration behavior of graduates in the U.S. and how that behavior affects career outcomes. The nuances of this relationship revolve around the graduates' decisions after completing their bachelor's degree. A graduate can migrate after completing their bachelor's, which leads to one of three outcomes: migrate for a job, stay in the same state for a job, or return to their home for a job. These actions are classified as Repeat Migrant, University Stayer, and Return Migrant, respectively. Additionally, a graduate can either remain at their alma mater then migrate later or never migrate from their alma mater. These decisions are noted as migrate for job and stay for job, respectively; these two decisions are Late Migrant and Non-Migrant. These classifications are crucial to analyze a migrant's self-selectivity. These classifications are tested in Kazakis and Faggian (2017) rudimentary hypothesis on whether a high propensity for migration is linked to higher salaries, on average. Kazakis and Faggian (2017) analyzes how graduates' annual salary is affected by the graduate's individual characteristics, education-related factors, and job-related traits. This analysis is the basis for comparing a graduates' expected returns from migrating versus not migrating. Consequently, Kazakis and Faggian (2017) discover both repeat migration and return migration have higher returns than non-migration at all, and late movers have the worst returns out of the five groups. Kazakis and Faggian (2017) approach inter-state migration in from a graduates' standpoint, there is more to inter-state migration then just education and jobs. Our research also factors in high-school education, which is another crucial to a graduate's attained success. A state that has a higher percentage of its population completing high school is far more desirable than a state with a low percentage. The implication is states with a high number of successful high school will create an increased demand for inter-state immigration, since a successful high school career is the basis for a successful college and job career. Additionally, our research analyzes how states are affected by inter-state migration, rather than just post-graduates' migration. Our research factors in number of Top 100 Universities in the U.S. as well as number of Fortune 500 companies. While these variables affect some graduates, these variables are extremely important to families that migrate. The parents will have some seniority and will migrate to a state with a better company while considering the future success for their children. Consequently, while Kazakis and Faggian (2017) approach inter-state migration in a nuanced manner through post-grad decisions, our research incorporates other factors to analyze inter-state migration more holistically.

Daniel Hummel (2015) looks at the relationship between state-level wellbeing and inter-state migration. In this approach Hummel looks considers the wellbeing of a state to be a potential explanation for why individuals might be incentivized to move into a state. Hummel suggests that as homogeneity on economic factors such as income and taxation rises across the US, the primary motivator for migration between states might switch from economic factors to non-economic wellbeing. One of the reasons Hummel gives for this change is due to a cultural shift in the US in which there has been more emphasis on quality of life factors than economic factors. These factors include, weather, public safety and a number of other elements that influence the subjective wellbeing of an individual. When controlling for economic and non-economic variables typically associated with in-migration it was found that there was a strong positive relationship between subjective wellbeing, in addition to the traditional objective wellbeing that looks primarily at economic measures of success such as per capita income or gross domestic product.

In the context of this paper we hope to explore if traditional factors such as income and education are still significant motivators for migration between states, despite findings that non-traditional factors such as subjective wellbeing and overall quality of life have increasingly become important motivators in an individual's decision to migrate between states. As such we will not be looking at the factors that Hummel explored in analyzing the relationship between wellbeing and migration but instead, we will be looking at factors that look at income (median household income and number of Fortune 500 companies) and education (high school graduation rate and number of top 100 universities). This paper hypothesizes that despite this trend toward non-traditional factors motivating migration, both income and education remain a significant motivator for migration between states.

3. Data

In order to answer if states with higher incomes also have more immigrants, certain variables are analyzed. Intuitively, the estimated number of immigrants per state is the dependent variable, while the natural log of a state's income is an independent variable. For this paper, a state's attractiveness to individuals will be determined in the model by the median household income. While Gross Domestic Product (GDP) can also be an indicator of a state's income, GDP can skew the results by giving too much influence to states with very wealthy individuals or states with which rely on exports to boost productivity. Whereas, the median household income is a better indication of a state's income, since this variable isn't easily skewed. Hence, median household income is a better variable for considering a state's income.

There are also other factors involved in determining the immigration that a state will receive; therefore, the paper will include as many of those variables as possible to avoid omitted variable bias and to use the other independent variables as control variables. One of the variables is the level of education of individuals in the state, which in this paper, is observed through the percentage of high school graduates in the labor force of a states. A high percentage of high school graduates suggests a state has successful and renowned high school programs, which would be important for families, since they want to give their children a better chance to succeed and would attract more immigration to the state. Naturally, a similar argument could be made for universities in a state. Therefore, the number of Top 100 Universities in a state will also be considered in the paper's analysis. Additionally, moving is economically sound, since in-state tuition is cheaper than out-state tuition. Certain faculty members are also enticed into immigrating to a different state for a renowned college, with the implication that a renowned college will guarantee better funding and equipment. Another factor to consider is the number of Fortune 500 companies in a state. People will naturally go to states with better companies. The reputation and success of these companies will boost immigration to their headquarter states and generate more income for that state. The final two variables are the cost of living for a state and the average annual temperature of a state. People naturally desire to maximize their profits; hence, people naturally gravitate towards states where the cost of living is lower, relative to other states. The cost of living index is the best method of comparing cost of living across states. The cost of living index is calculated based on the average costs of: groceries, housing, utilities, transportation, and health care. The resulting index is a holistic and accurate variable for measuring the cost of living. In addition to a state's cost of living, is the average annual temperature in that state. Since the average human body temperature is 98 °F, people find moderate temperatures more favorable than extreme temperatures. Additionally, there are fewer costs associated with moderate temperatures. Consequently, these are the variables analyzed in this paper.

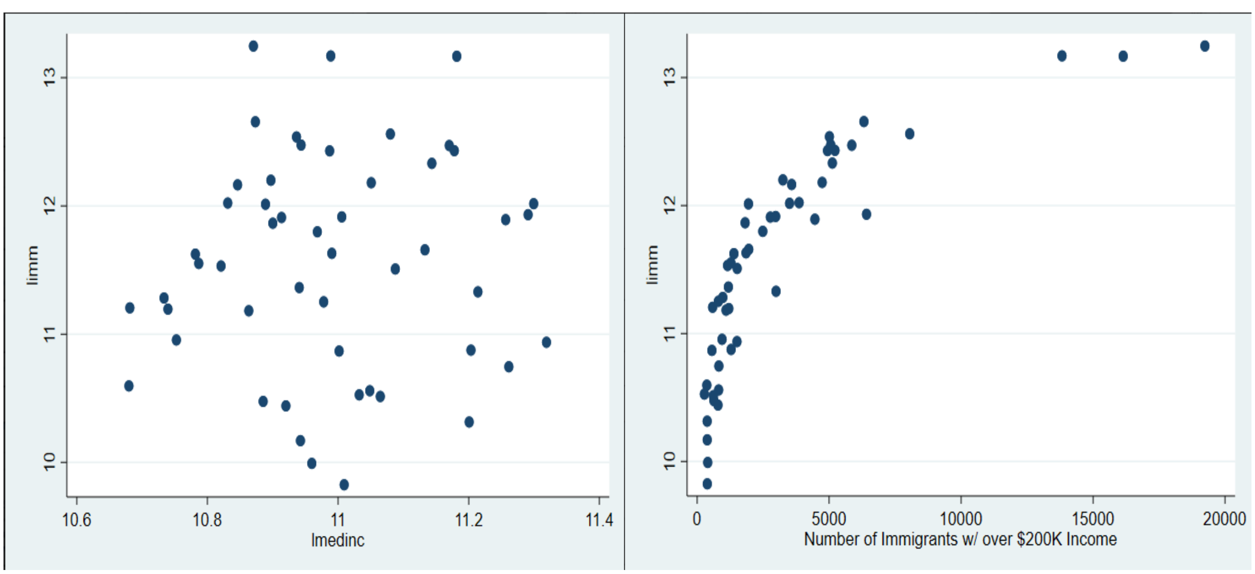
This paper's data originate from five distinct sources: census.gov for immigration numbers and median household income, irs.gov for the number of high-income immigrants, noaa.gov for the annual average temperature for each state, statistita.com for the cost of living index and Fortune 500 companies, and shanghairanking.com for the Top 100 Universities. Census.gov is the website that contains the census information for the United States of America, which includes inter-state migration for all 50 states. Additionally, the U.S. census lists all 50 states' median household income and average education percent estimates for high school. The Internal Revenue Services website, irs.gov, gives the annual reports of tax returns for the entire U.S. populace. These returns include those of high-income immigrants. Another source is noaa.gov, which is the website for the National Oceanic and Atmospheric Administration. The NOAA monitors various factors contributing to the U.S. climate, with an annual average temperature being one of the factors. Statistia.com simply lists where the 2017 Fortune 500 companies are based. While 2017 Fortune 500 companies' data can be found on other websites, statistia.com has everything compiled by state. Additionally, statistia.com calculates the cost of living index for each state based on the average costs of: groceries, housing, utilities, transportation, and health care. Finally, shanghairanking.com lists the 2017 Top 100 Universities in the U.S. Some unique traits shared by census.gov and shanghairanking.com are their longevity, reputation, and unbiased data. Census.gov reports data from the U.S. census which has been in use for over 200 years, while shanghairanking.com has been objectively ranking universities in the U.S. and around the world since 2003.

Variable	Interpretation	Year	Source
limm (dep.)	Natural log of immigration to a state	2017	US Census

lmedinc (ind.)	Natural log of state median income	2017	US Census
hsgrad (ind.)	Percentage of highschool graduates in the labor force	2017	US Census
fort (ind.)	Number of fortune 500 companies in the state	2017	statista.com
top (ind.)	Number of top 100 universities in the state	2017	shanghairanking.com
rich (ind.)	Number of immigrants with income over $200,000	2017	IRS
cost (ind.)	Cost of living index	2017	IRS
temp (ind.)	Annual Average temperature of the state	2017	NOAA

The first model which the paper will use for regression will be.

$$\square\square\square\square = \square_0 + \square_1\square\square\square\square\square\square\square\square + \square_2 h\square\square\square\square\square + \square_3\square\square\square\square + \square_4\square\square\square + \square_5\square\square\square h + \square_6\square\square\square\square + \square_7\square\square\square\square + \square(1)$$



The model will be justified based on its fulfillment of the Gauss-Markov assumptions for OLS estimates. This section will also clarify the possibility of biases from the way the data is being used and organized. The first assumption of linearity is fulfilled by the way the model has been established since all variables are all accounted for with linear slope relationships to *limm*. The second assumption to some extent is failed to be met by the model we put forth. The samples which were collected for the median household income data do reflect a random subset of the population of each state since the United States Census Bureau takes data in a random manner. The same can be stated of the sample data for the percentage of high school graduates within the states and the number of college graduates within the states since that information, too, is collected from the United States Census Bureau. However, where this model begins to fail to meet the Gauss-Markov assumptions involves the information regarding the number of Fortune 500 companies within a state and the number of top 100 universities within a state. These two data sets are not obtained through random sampling; they are instead samples of companies and universities which meet the criteria set forth by *Fortune* magazine in the case of the number of Fortune 500 companies in a state and by the Academic Ranking of World Universities in the case of the top 100 universities in a state. Despite the failure of the model to fulfill this Gauss-Markov assumption, excluding these two variables would hurt the model more than help it. Since the number of Fortune 500 companies and the number of top 100 universities that are in a state would also influence in the median income and the percentage of high school and college graduates. Furthermore, the inclusion of a wide variety of economic factors along with the non-economic factor of average temperature helps account for any potential omitted variable bias. The model also fulfills that third of exogeneity Gauss-Markov assumption since there is no perfect collinearity between the variables; however, it is worth noting the high correlation between *fort*, *top*, and *rich*, which hovered between 70 and 80 percent. After making these bias assumptions, the paper will also assume homoscedasticity and assume that the results for this model will be the best linear estimate that will exist for the model. Now that the paper has stated the assumptions that it is making and the assumptions that it is not, the paper will now proceed to show the results from the STATA regression analysis.

4. Results

The first model which this paper tested was a simple linear regression between *limm* and *lmedinc* in the model:

$$\square\square\square\square = \square_0 + \square_1\square\square\square\square\square\square\square + \square \quad (2)$$

The regression for this model gave the results:

Regression Between log(Immigrants) and log(Median Household Income)	
Intercept (β_0)	8.0 (7.94)
Slope (β_1)	0.3 (0.72)
$\square^2$	0.004
# of Observations	51

Note: The quantities in parentheses below the estimates are the standard errors. *Significant at 10%, **5%, ***1%

The model can be written as:

$$\hat{y} = 8.0 + 0.3\square_1 (3)$$

This model gives the impression that a 1 percent increase in *lmedinc* would correspond to a 0.3% increase in *limm*, which would indicate that the immigration to and from a state is only impacted slightly by the median income of the state. This small effect may be due to the lack of other controlling variables in the model. Furthermore, the small R^2 of the model indicates a lack of reliability for the model and that only 0.4 percent of the variation in the sample is explained by the model. A t-test was conducted to assess the statistical significance in this model. It was found that lmedinc has a t-value of 0.45 and the intercept had a t-value of 1.01. Both are below the critical value of 1.68 found at a 10% level of significance for a two-tail test with 50 degrees of freedom. As such neither lmedinc or the intercept are statistically significant. Though this relationship is positive it is weak and justifies the inclusion of the other controlling variables.

The multiple linear regression is given as model (1):

$$\square\square\square\square = \square_0 + \square_1\square\square\square\square\square\square\square\square + \square_2 h\square\square\square\square\square + \square_3\square\square\square\square + \square_4\square\square\square + \square_5\square\square\square h + \square_6\square\square\square\square + \square_7\square\square\square\square + \square (1)$$

The regression for this model gave the results:

Multiple Linear Regression on log(Immigration)	
Intercept(β_0)	11.5** (5.35)
log(Median Household Income) (β_1)	-0.23 (0.55)
High School Graduate Rate(β_2)	0.01 (0.04)
Number of Fortune 500 Companies(β_3)	0.02* (0.01)
Number of Top 100 Universities(β_4)	0.02 (0.06)
Number of immigrants with income over \$200,000(β_5)	0.0001*** (0.00)
Cost of living index(β_6)	-0.001** (0.00)
Annual Average temperature of the state(β_7)	0.02* (0.01)
$\square^2$	0.7220
# of Observations	51

Note: The quantities in parentheses below the estimates are the standard errors. *Significant at 10%, **5%, ***1%

And can be written as:

$$\hat{y} = 11.5 - 0.23\square_1 + 0.01\square_2 + 0.02\square_3 + 0.02\square_4 + 0.0001\square_5 - 0.001\square_6 + 0.02\square_7 (4)$$

This model firstly contradicts the results of model (3), the simple regression model. Whereas in (3) there is a weak positive correlation between *lmedinc* and *limm*, there is now a weak negative correlation, which is implying that a 1 percent increase in *lmedinc* would result in a 0.23 percent decrease in *limm*. The inclusion of the other variables, however, proved more effective in giving a greater scope into the factors that affect immigration. As expected *hsgrad*, *fort*, *top*, *rich*, and *temp* all had positive effects on *limm*. Also expects is the negative relation between *cost* and *limm*.

Analyzing each variable independently, a better understanding of their effects can be understood. *Hsgrad* has a relatively steady effect on *limm*; as a 1 percent increase in *hsgrad* will cause a 1 percent increase in *limm*. *Fort*, *top*, and *temp* have greater effects on *limm* since they are in terms of percentages. An increase of 1 in any of these variables will result in *limm* increasing by 2 percent since they are in a level-log relation with *limm*. To clarify, if a state gets one more top 100 university, if a state gets one more Fortune 500 company or if a state's average temperature increases by 1 degree its immigration will increase by 2 percent. However, the effect of *temp* would probably not continue indefinitely since an average temperature consistently in the 200-degree range would not encourage immigrants to come to a state. *Rich* has the smallest impact on *limm* since an increase of 1 immigrant will increase immigration by 0.01 percent in a state, which indicates that rich immigrants have a weak pull effect on the actions of other immigrants. *Cost* also has a small effect; however, it is negative. For every increase in the cost of living index by 1, the immigration to a state decreases by 0.1 percent. In model 4 and model 3 the constant term is of little importance to the paper since there are no reasonable situations in which there would be a state with a median income of 0 or a cost of living of 0.

Though these effects are all relatively small on their own, the model proposes a strong relationship between them with an R^2 that implies that 72.20 percent of the sample's variation is explained by the model. However, before a conclusion can be drawn the paper will have to look into statistical interpretation of the results of the regression.

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
lmedinc	-.23	.56	-0.42	0.678	-1.35	.89
hsgrad	.01	.04	0.36	0.719	-.06	.09
fort	.02	.01	1.76	0.085	-.00	.04
top	.02	.06	0.34	0.735	-.09	.13
rich	.0001	.00	3.32	0.002	.00004	.0001
cost	-.001	.00	-2.16	0.036	-.002	-.00007
temp	.02	.01	1.71	0.095	-.004	.05
_cons	11.47	5.35	2.14	0.038	.67	22.27

Based on the STATA regression there are various conclusions that can be drawn from the results of the model 1 and model 4. The low t-scores for most of the variables excluding *rich*, and *cost* indicate that at most levels of certainty the coefficient of the variables along with their standard errors will fail to disprove the hypothesis that the coefficients for the variables should be zero. In other words, the STATA regression indicates that the independent variables have a relatively negligible effect on the dependent. Using the P-values from the table it can be interpreted that only *rich* has a significant impact on *limm* at even the 1 percent level. *Cost* comes in second of importance with P-score of 0.036, which indicates that at the 3.6 percent of certainty and any percent high including 5 percent and 10 percent *cost* has an effect on *limm*. The next two important variables are *fort* and *temp* with P-scores of 0.085 and 0.095 respectively which indicates that the variables are significant at the 8.5 percent and above certainty or the 9.5 percent certainty and above.

Lastly the confidence interval help solidify the interpretation of the P-value. The range which the chart gives indicates with 95 percent certainty all the possible values, which the coefficient of that variable could have. To ensure that the variable has significance, the number 0 must not be in this range because if the coefficient was 0 the variable would have no impact on *limm*. Using the confidence interval reveals that only *rich* and *cost* do not have a 0 in their confidence intervals. This indicates that with 95 percent certainty, *rich* and *cost* do have a significant impact on *limm* while the other variables do not.

5. Extensions

Using the results from the first regression, further analysis can be conducted to improve the model and achieve a more adequate and representative model. Since *rich* and *cost* both have significance at the 5 percent level, they will be left alone for further analysis for now.

The first analysis will be a F-test involving *fort* and *top*. Since the two variables had high collinearity there is a possibility that together they will be statistically significant compared to independently. To perform this F-test the paper will compare model 1 with this model and with the following assumptions:

$$□□□□ = □_0 + □_1□□□□□□□ + □_2h□□□□□ + □_5□□□h + □_6□□□□ + □_7□□□□ + □ \quad (5)$$

$$□_0 : □_{□□□□} = □_{□□□} = 0$$

$$□_1 : □_{□□□□} \neq □_{□□□} \neq 0$$

$$□ = 2 \quad □□ = □ - 1 - □ = 43 \quad □ = 51 \quad □ = 7 \quad R^2_{UR} = 0.7220 \quad R^2_R = 0.6834$$

$$□ = \frac{(□^2_{□□} - □^2_{□})/□}{(1 - □^2_{□□})/□□} = \frac{(0.7220 - 0.6834)/2}{(1 - 0.7220)/43} = 2.99$$

Then by comparing the F-score of the *fort* and *top* combined to the F-score at a 5 percent certainty of $F_{7, 51,}$ which is between 2.17 and 2.25, it becomes clear that the the F-score of *fort* and *top* is greater. The result rejects the null hypothesis and indicates that at a 5 percent confidence level, *fort* and *top* have joint significance, this is important in developing a future model by showing that the model should have only *fort* or only *top* to stop collinearity to decrease the significance of the results.

Though there was collinearity between *rich* and *fort* as well as between *rich* and *top*, there will be no f-test analysis for *rich* and the other variables, and they will not be joined in the next model due the significance that *rich* has as a variable on its own. Still looking forward to the new model, *hsgrad* will be dropped from the model due to its low P-value and its lack of reasonable collinearity with other variables. Though *hsgrad* and *temp* do have a negative collinearity of -0.69, this is likely not an accurate reflection of the factors that affect *hsgrad*, so it will be dropped.

After dropping the variables with a significance below 10 percent and then joining *top* and *fort* to just *fort*, the new regression model is:

$$□□□□ = □_0 + □_1□□□□□□□ + □_2□□□□ + □_3□□□h + □_4□□□□ + □_5□□□□ + □ \quad (6)$$

The model is retaining the *lmedinc* since the purpose of this analysis is still to view the effect of *lmedinc* on *limm*. The regressing this model results in the following coefficients and standard errors.

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]

lmedinc	-.090	.44	-0.20	0.840	-0.986	.806
fort	.018	.007	2.53	0.015	.004	.034
rich	.0001	.000	3.75	0.000	.00004	.0002
cost	-.001	.0005	-2.18	0.035	-.0019	-.00007
temp	.018	.010	1.94	0.06	-.0007	.038
_cons	11.19	5.07	2.21	0.032	.979	21.41

These results help solidify a more useful and representative model than model 4. The similarities between model 4 and model 6 include the negative correlation between *limm* and *lmedinc* and the same positive and negative correlations for the other variables too. Model 6 does have a stronger case for accurately depicting the variables that affect *limm* as all of the variables with the exception of *lmedinc* have are significant at least at the 10 percent level and only *temp* isn't significant at the 5 percent level. However, in model 6 just like model 4 there is little evidence for a significant effect of *lmedinc* on *limm*.

6. Conclusion

Unfortunately for the hypothesis paper, median household income, unintuitively, has a negative relation to immigration; however, this correlation has failed to be statistically significant, so it is more apt to say that the findings of this study are inconclusive at best. This result was unexpected based on the evidence that other prior papers. The relation between immigrants with income over $200,000 and immigration was expected because those individuals factor directly into the measure of immigrants for a state. The subsequent worthwhile measures are the cost of living in the state and the number of Fortune 500 companies in a state, which had a 5% statistical significance, along with the temperature which was only statistically significant at the 10% level. The significance of temperature was also partially expected based on the research performed prior to this paper.

Looking forward, further research would need to focus on the impact of cost of living on state immigration instead of median household income; furthermore, there would also need to be research done on the reason for lack of correlation between median household income and state immigration. Also useful would be research including other factors that are related to cost of living in order to more accurately reflect the true effect of cost of living without omitted variable bias.

References:

Cebula, R & Alexander G. "Determinants of Net Interstate Migration," *The Journal of Regional Analysis and Policy*, (2015) 36:2. Retrieved from https://ageconsearch.umn.edu/record/132323/

Frey, W.H. "Immigration, Welfare Magnets and the Geography of Child Poverty in the United States," *Population and Environment* (1997) 19: 53. Retrieved from https://doi.org/10.1023/A:1024697515511

Kazakis, P. & Faggian, A. "Mobility, education and labor market outcomes for U.S. graduates: Is selectivity important?," The Annals of Regional Science (2017) 59: 731. Retrieved from https://doi.org/10.1007/s00168-016-0773-6

Index

Collinearity Test:

```
. corr lmedinc hsgrad fort top rich cost temp
(obs=51)
```

	lmedinc	hsgrad	fort	top	rich	cost	temp
lmedinc	1.0000						
hsgrad	0.4853	1.0000					
fort	0.2302	-0.2970	1.0000				
top	0.3157	-0.2821	0.8400	1.0000			
rich	0.1696	-0.3575	0.7231	0.7542	1.0000		
cost	0.0798	0.1921	-0.0820	-0.0608	-0.0813	1.0000	
temp	-0.2989	-0.6967	0.1367	0.1929	0.4030	-0.1535	1.0000

Simple Linear STATA Regression:

```
. regress limm lmedinc
```

Source	SS	df	MS			
				Number of obs	=	51
				F(1, 49)	=	0.20
Model	.145405581	1	.145405581	Prob > F	=	0.6573
Residual	35.7664082	49	.729926699	R-squared	=	0.0040
				Adj R-squared	=	-0.0163
Total	35.9118138	50	.718236276	Root MSE	=	.85436

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lmedinc	.3222141	.7219276	0.45	0.657	-1.128554	1.772982
_cons	8.017842	7.935484	1.01	0.317	-7.929111	23.96479

Multiple Regression STATA:

```
. reg limm lmedinc hsgrad fort top rich cost temp
```

Source	SS	df	MS			
				Number of obs	=	51
				F(7, 43)	=	15.95
Model	25.9275071	7	3.70392958	Prob > F	=	0.0000
Residual	9.98430673	43	.23219318	R-squared	=	0.7220
				Adj R-squared	=	0.6767
Total	35.9118138	50	.718236276	Root MSE	=	.48186

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lmedinc	-.2317278	.5550155	-0.42	0.678	-1.351023	.8875677
hsgrad	.0130519	.0360134	0.36	0.719	-.059576	.0856798
fort	.0177047	.0100476	1.76	0.085	-.0025581	.0379676
top	.0188633	.0552617	0.34	0.735	-.0925825	.1303091
rich	.0001018	.0000307	3.32	0.002	.0000399	.0001637
cost	-.0010222	.0004732	-2.16	0.036	-.0019765	-.0000679
temp	.0208959	.0122234	1.71	0.095	-.0037549	.0455467
_cons	11.47233	5.353937	2.14	0.038	.6750913	22.26958

Restricted Model Regression:

Source	SS	df	MS
Model	24.5419487	5	4.90838975
Residual	11.3698651	45	.252663669
Total	35.9118138	50	.718236276

Number of obs	=	51
F(5, 45)	=	19.43
Prob > F	=	0.0000
R-squared	=	0.6834
Adj R-squared	=	0.6482
Root MSE	=	.50266

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lmedinc	.1373463	.5357781	0.26	0.799	-.9417661	1.216459
hsgrad	-.0203992	.0345031	-0.59	0.557	-.089892	.0490936
rich	.0001527	.0000221	6.91	0.000	.0001082	.0001972
cost	-.0010566	.0004931	-2.14	0.038	-.0020498	-.0000633
temp	.0096435	.0118101	0.82	0.418	-.0141433	.0334302
_cons	10.96518	5.423208	2.02	0.049	.042281	21.88808

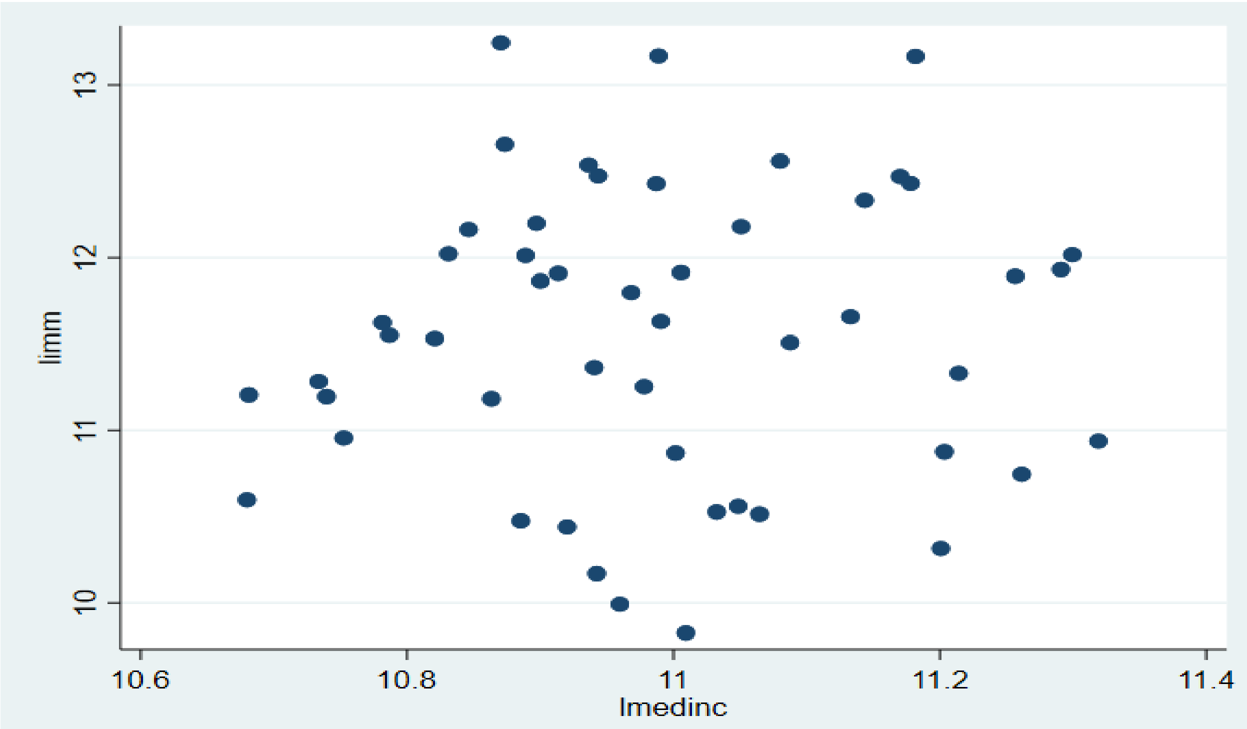
Model 6 Regression:

Source	SS	df	MS
Model	25.8779247	5	5.17558494
Residual	10.0338891	45	.222975314
Total	35.9118138	50	.718236276

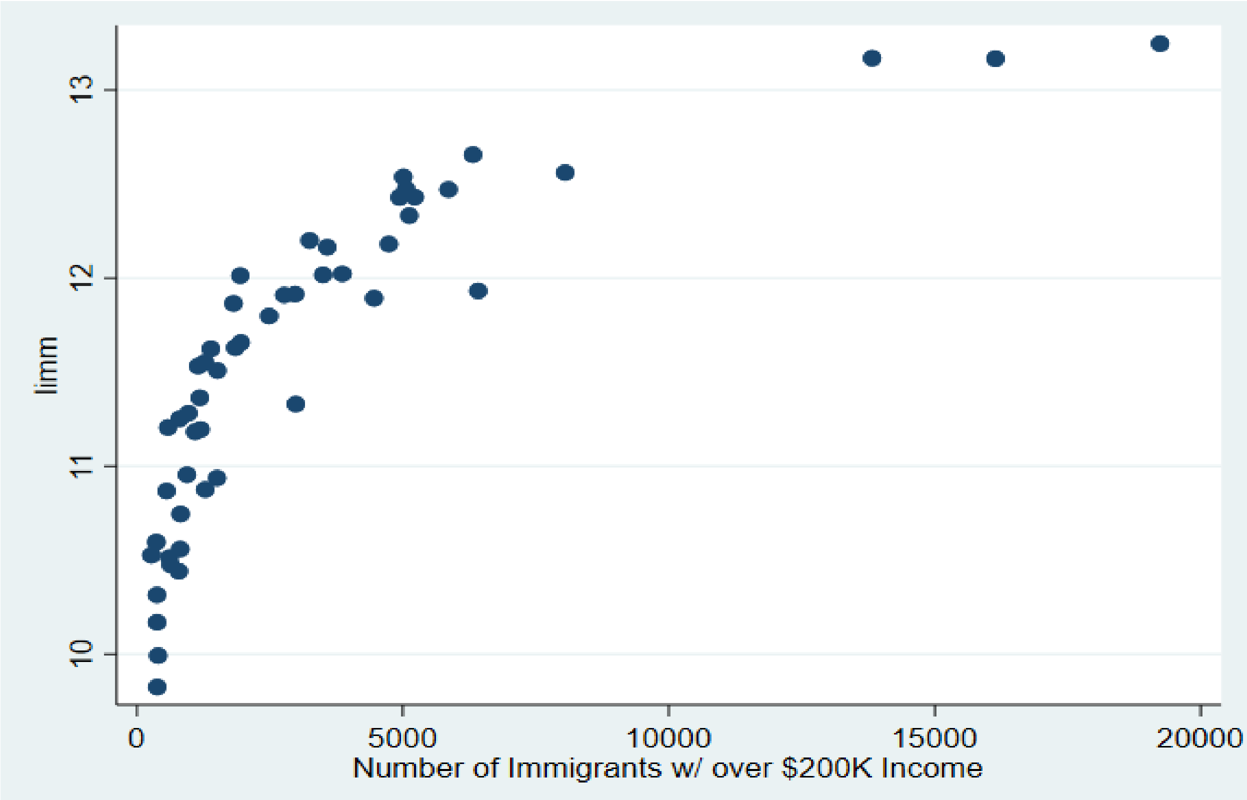
Number of obs	=	51
F(5, 45)	=	23.21
Prob > F	=	0.0000
R-squared	=	0.7206
Adj R-squared	=	0.6896
Root MSE	=	.4722

limm	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lmedinc	-.0900791	.4448235	-0.20	0.840	-.9859995	.8058414
fort	.0187819	.0074314	2.53	0.015	.0038144	.0337494
rich	.0001059	.0000282	3.75	0.000	.0000491	.0001627
cost	-.0010045	.0004616	-2.18	0.035	-.0019344	-.0000747
temp	.018408	.0095114	1.94	0.059	-.0007489	.037565
_cons	11.19233	5.070863	2.21	0.032	.9790866	21.40557

Scatter-plot of *limm vs lmedinc*:



Scatter-plot of *limm vs rich*:



Scatter-plot of *limm vs temp:*