Patrick Burns
12/12/19

# The Effect of Marriage Rates on Household Income

## Introduction:

The dataset that I chose to analyze comprised of census and demographic data across the United States. To make my data easier to interpret and work with, I narrowed it down to records in Illinois, Iowa, and Wisconsin because I have a fair amount of background knowledge for these states and they are all close together geographically. The dataset was very complete with 80 different attributes to look at, some more useful than others. My data really did not require much attention to clean, the only thing I really did was remove a mostly unused column, "BLOCKID", and deal with a few missing values. Upon looking more at what attributes I had to work with, I decided I wanted to use household income as my response variable because there are many different attributes that might be correlated with this. Further, I had to decide whether I wanted to look at household or family income. I decided household income would be better because it will include people who are living alone or not with family members to get a more holistic idea of income in various areas. Similarly, I chose to look at median household income rather than mean because median is resistant to outliers, and again could provide me with a better idea of where a typical household stands in each city. To determine what I would use as my explanatory variable, I created some visualizations to give me a general idea as to whether two variables were related. The visualizations that showed signs of potential relationships were marriage, population, and median rent. The relationship that appeared strongest was marriage vs household income, which became even more defined once I used small multiples to break that down between the three states. Thus, I decided that the focus of my project would be looking into the relationship between marriage rates and median household income in different areas. The overarching research questions I will be investigating are (1) What is the relationship between marriage rates and median household income, and (2) Is there a significant difference in this relationship between different states?

## Literary Review:

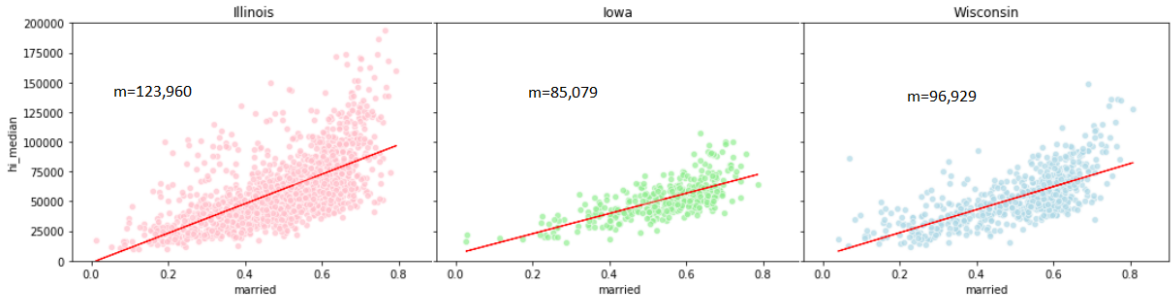
I intentionally chose to look at Illinois, Iowa, and Wisconsin because I have a fair amount of background knowledge about each of the states to start with, and because they are all located near each other and are easy to compare without having to concern ourselves with regional differences. One thing that separates Illinois that I wanted to think about when researching was the effect of the Chicagoland area. All three of the states I am looking at have a lot of rural areas that make up most of the state, however Illinois is different from the others because of the

unique, large, metropolitan area surrounding Chicago. While Chicago itself is by far the largest city across all three states, this will make up just a few leverage points because the data is based on location. What will have a much more significant effect is the "Chicago suburbs", because most of Northwestern Illinois is made up of super large, very wealthy, urban areas. This will likely to show up in the data as many points which will reflect areas where median household income is much higher than the average across the rest of the states. As such, many of my articles were focused on looking into this suburban effect and potential differences across the three states. Once I began reading some journals, I learned that other studies have found that marriage rates increase as a result of higher income. This was not necessarily surprising to me, although it is interesting given that I structured my analysis with marriage rates being the explanatory variable. This shouldn't be a problem for me since I am using a linear regression to show correlation, not causation. I also found that individuals who earn more money can go either way, in that these people will get more marriage offers but also being self-reliant is a viable option for them so this may neutralize this effect to some extent. Another interesting factor to consider is that the middle-class is shrinking rapidly in large cities. In Chicago there is already almost no middle-class, just upper-class and lower-class households. This could affect my analysis because the upper-class neighborhoods will likely show up way above any predictive trend line I create, while the lower class will fall below. Following that idea, I found that income inequality is especially great in suburban areas, and that in metropolitan areas it follows a donut-like pattern, with inequality growing as you move outwards. This would suggest that Chicago's surrounding suburbs would be just as bad if not worse than the city itself in terms of inequality. One study also suggested that inequality increases as you move up the "urban hierarchy" which would mean that I can expect the large suburbs and cities found in Illinois to show much more variability in income data.

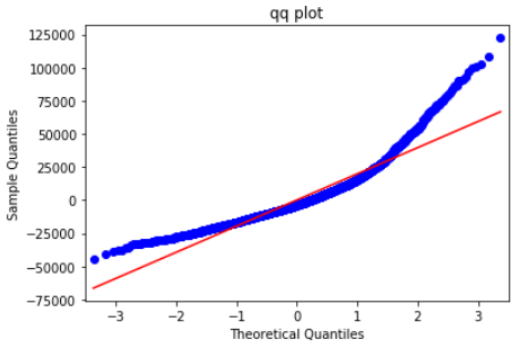
**Method:**

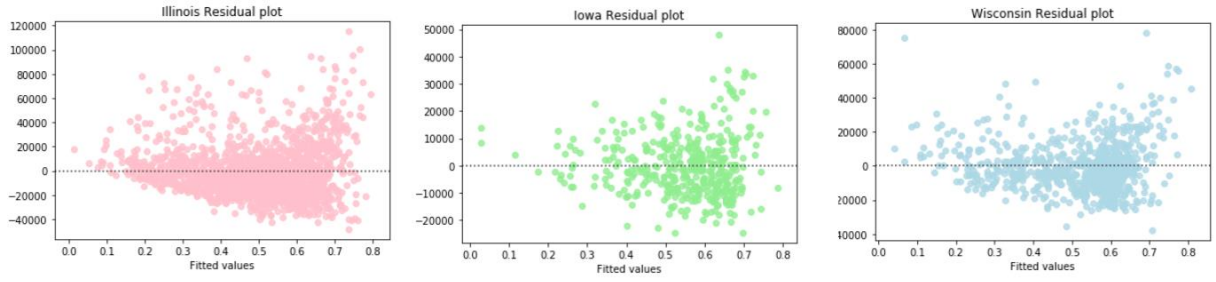
The first thing I did to manipulate my data was very minor cleaning. I began by narrowing my dataset down to just the three states that I will be looking at: Illinois, Iowa, and Wisconsin. Next, I completely removed the "BLOCKID" attribute because it was blank for almost all of the entries, and it was not something I needed for analysis anyways. Finally, I had to deal with missing income data. In total, only 34 out of 2,740 entries were missing data, so I decided to print the values that I would be dropping if I chose to ignore them. I found that there were a few small "clumps" of missing data from similar locations which might suggest that simply dropping them might systematically leave out relevant data. In the grand scheme of the data I decided this was not enough to significantly skew my results, especially because it was so few entries. The missing data also came from all three states, and ten different counties. So I continued on and decided to just drop the rows that were missing income data. Now that my data had been cleaned, I moved into analysis. I started by creating scatterplots of marriage versus median household income to get a visual representation of the relationship the two attributes

might have. I displayed the scatterplot of all three states combined first, followed by the three individual states side by side for easy comparison. The first plot with all three was not particularly useful. It suggested a positive trend, but it still looked a lot like a large blob. Comparing that to the scatterplots of each individual state was a lot more helpful. I could see immediately some difference between the shape of the data between states. Iowa and Wisconsin were both very similar, while Illinois had many more points that reflected a very high household income and generally appeared as though a regression would have a steeper slope. I then calculated and plotted the regression lines for each of the three states. I put these three graphs on the same scale and side by side so I could easily compare them visually, as displayed below.




This confirmed my initial idea that Illinois would have a significantly steeper slope than the other two states. I wanted to check if it was appropriate to use a linear model, so I checked both heteroskedasticity and normality for all three plots using the Breusch-Pagan and Shapiro tests. For all of the tests I performed, the p-values were small so I was not able to reject any of the null hypotheses. I then created a qqplot to better assess the lack of normality in my data. This plot appeared to be tail-heavy, especially on the right side. This suggested some right skew, which was not surprising. I created residual plots as well so I could look for potential leverage points. Illinois appeared to have many more than the other states as expected, but now I could see that almost all the leverage points fell above, not below the horizontal line. The final analysis I did was testing to see if the difference




in slope for Illinois was statistically significant. I attempted to do this with a Kruskal Wallis test

first because I was thinking I have 3 different states to look at with unequal variance and sample sizes, however I realized that this was not actually testing for what I wanted to find. I was looking to evaluate the difference in slope between Illinois versus the rest of the data, so a two-sample t-test was more appropriate. As such I found summary statistics for Illinois and for Iowa and Wisconsin combined, and plugged these values into the test.

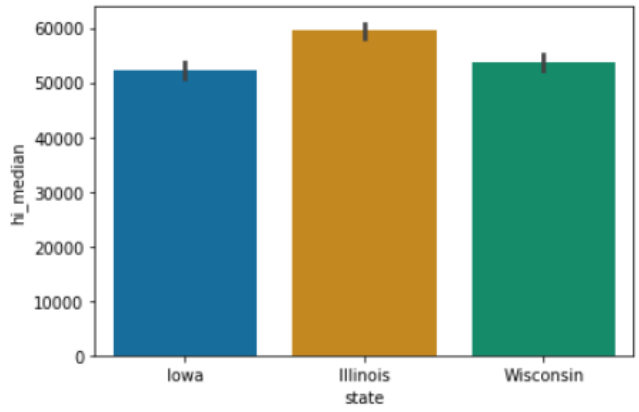
**Results and Analysis:**

The first main takeaway I had from my analysis was that there was some linear relationship between marriage rates and average household income. The scatterplots suggested this, and then I used a linear regression for modeling because I wanted to find a predictive relationship between marriage rates and household income. The data was quantitative and did not easily fall into groups so a logistic regression would not have fit well. It is important to take this model with a grain of salt because it did fail the necessary assumptions. All three state models were heteroskedastic and not normal when I conducted tests on them, which tells me that my linear model could be flawed. The residual plots I created also showed many potential leverage points above the horizontal line in all three models, with Illinois having the most. These points certainly affected the slope of my regression lines, most likely pulling the slope upwards as that is where most of the leverage points were. When comparing the actual regression equations, Illinois stood out having a higher slope than Iowa and Wisconsin by almost 20,000. Further, it

| Illinois | Iowa | Wisconsin |
|---|---|---|
| Intercept: [-1643.32151146] | Intercept: [5676.66006947] | Intercept: [4167.54828904] |
| Slope: [[123959.79538389]] | Slope: [[85079.27628611]] | Slope: [[96929.71251465]] |

stood out in that both Wisconsin and Iowa had similar intercepts, which fell into the 4000-6000 range, while Illinois had a negative intercept of -1643. This could be a result of the excessive number of leverage points in Illinois, and again points towards my research idea that state does affect the relationship between marriage rates and household income. I also created a bar graph showing the average median household income in each state which showed Illinois to be higher than both Iowa and Wisconsin. This was interesting since Illinois had the highest median income, meaning if the three states all had a similar relationship between marriage and income we could expect that the intercept

```
IL: 59520.728198599616
IA: 52352.108695652176
WI: 53793.133148404995
```



would actually be highest in Illinois, and that all three slopes would be closer together. Instead, the slope in Illinois was much greater, leaving us with a negative intercept. While I could see a numerical and visual difference in slope from the graph and equations, the output from my two-sample t-test was able to confirm the relevance of this difference. The resulting p-value was 0.0, which suggests that Illinois' difference in slope is statistically significant and should not be attributed to random error.

**Discussion:**

As a whole, the tests that I conducted confirmed my initial predictions. Regarding my first research question, I found that there was a positive correlation between marriage rates and household income. This suggests that people who live in areas where a higher proportion of the population is married will earn a higher income on average. This might be because marriage is associated with stability in a person's life, so as people have their lives more put together, they will get married more. This is backed by some of the research I found, which shows that people with more financial stability tend to get more marriage offers which would raise marriage rates. This is working opposite of how I structured the relationship in that income is acting as the explanatory variable. I believe marriage rates might raise average income because people who have stable relationships and are supporting a family would have more incentive to work harder and earn a higher income for those dependent on them. My analysis also answered my second research question, showing that this relationship was different in Illinois. It seems that the effect was exacerbated in Illinois. At first this conclusion did not make much sense to me, I could not think of a logical reason that marriage rates would affect income differently in Illinois so I attributed the difference primarily to the great number of high leverage points that come from the Chicago suburbs. My research helped to explain this trend, specifically the findings that income inequality increases as you move up the "urban hierarchy". Because of the Chicagoland area, Illinois has many rich, large suburbs that Wisconsin and Iowa do not. This would mean that we can expect both average income and income inequality to be higher in Illinois. Further, the Chicago suburbs and other urban areas like it are often regarded as good places to raise a family, so we can expect higher marriage rates and income in those same areas.

While my research can explain the leverage points, the Illinois model should still be taken with a grain of salt because the entire regression equation is impacted by these suburbs. As such, my equation will systematically over-predict more than it under-predicts. The other reason we should be cautious with my findings is because of the heteroskedasticity and lack of normality of the data. I was not surprised to find this because I am looking at income data. Much of my research suggested that I would find a lot of income inequality. Simply based off general knowledge of wealth distribution, we can expect that the top 1% will have a drastically higher income than the rest of the population. This can be seen in the qqplot I created; the right side spiked up above the line suggesting a right skew to the data. If we needed a very accurate model,

my linear regression would not be sufficient since both heteroskedasticity and lack of normality tell us a standard linear regression is not appropriate. It is, however, the most fitting model that I know how to create at this point. If this data was to be analyzed more professionally, they would likely use a different type of regression, or at least find some way to account for leverage points better than I could. Finally, we should be cautious with accepting the results of my t-test. While I do have some statistical background, I am far from a professional. I believe that this test would be much more meaningful if we were testing Illinois versus the Midwest as a whole, rather than just Wisconsin and Iowa. Testing against a larger area would better reflect what a standard relationship between marriage and income looks like and would give the test more meaning. Since I only tested against two other states it was not very surprising or meaningful to find that the slope of the equation was significantly different.

To make use of my analysis, we would need to do further research. My conclusion suggests that people who live in places with high marriage rates will earn more money, so the next logical step would be to investigate why this might be. We should look at particular areas and analyze job opportunities there. Are there better opportunities in these rich suburban areas? Or have married people moved up the ladder in their jobs to earn the extra money? Some of my research showed that high-paying job opportunities are shifting towards higher-populated areas, so for college graduates, living in a rural area is usually not a logical decision. One other way we could build on this analysis is looking at areas over time. How do marriage rates and median income vary with time? If we found they both rise and fall together, this would be consistent with my findings, but otherwise we might have some evidence that suggests income is more heavily influenced by other factors.

Going forward, if we wanted to take my analysis further we should look at what other variables might play into average household income. It would be interesting to test for significance with each of these different variables, and perhaps find some way to make a more complex model that can account for multiple factors. Another way to build on my analysis would be to compare my findings to similar analyses in a different region of the United States, or even another country. Perhaps in some areas marriage rates and household income would have an inverse relationship, and then further research could be conducted to see why this might be.

**Conclusion:**

While my methods may be flawed, my analysis does provide some evidence that there is a positive linear relationship between marriage rates and median household income across Illinois, Iowa, and Wisconsin. I found that this relationship had a steeper slope in Illinois, but was not necessarily as strong of a relationship given the extra variability in the data. These results suggest that people who live in areas where a greater percentage of the population are married earn a higher income, and that this effect may be even greater in Illinois than it is in Wisconsin and Iowa.

References:

Burgess, S., Propper, C. & Aassve, A. The role of income in marriage and divorce transitions among young Americans. J Popul Econ 16, 455–475 (2003) doi:10.1007/s00148-003-0124-7

Cook, L. (2015, October 26). For Richer, Not Poorer: Marriage and the Growing Class Divide. Retrieved December 11, 2019, from https://www.usnews.com/news/blogs/data-mine/2015/10/26/marriage-and-the-growing-class-divide.

Gibbs, R. M. (1995). Going away to college and wider urban job opportunities take highly educated youth away from rural areas. *Rural Development Perspectives*, *10*, 35-44.

Long, J., Rasmussen, D., & Haworth, C. (1977). Income Inequality and City Size. *The Review of Economics and Statistics*, *59*(2), 244-246. doi:10.2307/1928824

Lutton, Linda. "The Middle Class Is Shrinking Everywhere - In Chicago It's Almost Gone." WBEZ, WBEZ, 20 Feb. 2019, https://www.wbez.org/shows/wbez-news/the-middle-class-is-shrinking-everywhere-in-chicago-its-almost-gone/e63cb407-5d1e-41b1-9124-a717d4fb1b0b.

Peters, D.J. (2012), Income Inequality across Micro and Meso Geographic Scales in the Midwestern United States, 1979–20091. Rural Sociology, 77: 171-202. doi:10.1111/j.1549-0831.2012.00077.x

W. Levernier, M.D. Partridge, D.S. Rickman. Differences in metropolitan and nonmetropolitan U.S. Family income inequality: A cross-county comparison. Journal of Urban Economics, 44 (2) (1998), pp. 272-290

Worthington, R. (2018, August 29). REPORT TIES SUBURBAN SPRAWL TO AFFLUENCE. Retrieved December 11, 2019, from https://www.chicagotribune.com/news/ct-xpm-1998-11-03-9811030111-story.html.