-3 = 30/33

**Income Inequality in the Midwest**

Connor Boone, Rakshanda Khan, Brandon Welker

**Introduction:**

The topic of income inequality is one that has captured the attention and inflamed the passions of the general public throughout history, especially in recent years. The Occupy Wall Street movement of 2011 demonstrated the rage that many Americans felt about an economic system perceived as rigged to benefit the top 1% of wealthy Americans. Great and growing differences in earnings and wealth between members of society has the ability to incite resentment, fear, and premature death. Indeed, the case was made in the American Journal of Public Health by McDonough, et al, that low income and income instability can be linked to higher mortality rates (2011).

Policymakers have high incentive to pay close attention to trends of and attitudes about extreme income disparities, as the consequences of rampant inequality can be a danger to society, whether locally, nationally, or globally. (need source, maybe quote GINI paper) This paper will focus on causes of income inequality at the local level in the states of Indiana, Kentucky and Ohio. This analysis examines the effects to which certain social statistics that are postulated to drive variance in economic equality affect GINI at a county level within the scope of the model.

The GINI coefficient, developed by Italian statistician Corrado Gini in 1912 neat, will function as the dependent variable. The coefficient, which ranges from zero (perfect equality in within a population) and one (all wealth or income in a population held by one individual). Using our county-level data gleaned from the U.S. Census bureau and other sources, we will attempt to demonstrate the relationship between economic growth and income inequality. The Lorenz Curve is a graphical representation of the cumulative income earned by a given percentile of people sorted by income from lowest to highest. The GINI coefficient describes the degree of difference from the cumulative distribution of income in the Lorenz Curve from that of perfect equality of income, or linear equality. Using our county-level data gleaned from the U.S. Census bureau and other sources, we will attempt to demonstrate the relationship between economic growth and income inequality.

In the United States, the study of income inequality can be traced back to the early 20th century. Early works by Edwin Cannan’s “Division of Income” (1905) and H.J. Davenport’s “Value and Distribution” (1908) mark the earliest mentions of the issue of income distribution. More recently, the conversation has shifted towards examining the economic, political, and societal forces driving the mostly consistent growth in inequality. Nationally, a study out of UC Berkeley (Saez, 2016) tracked the rise in income share of the top 10% and top 1% of incomes over nearly the last century. As can be seen in Figure 1 below, the top 10% of earners in the U.S. have never accounted for less than 30% of all income nationally. Further, the top earning decile recently accumulated as much as half of all income.

Levernier, Rickman, and Partridge (1995) examined economic data for the contiguous 48 states in the U.S. and studied inequality trends at a regional level. Key findings of their research were that some notable causes of inequality increase during the 1980’s were due to heightened international immigration and the rise of households headed by a single female. Additionally, their research supported arguments that factors such as greater high school attainment, labor force participation, goods-producing employment share, and transfer payments aided in reducing or mitigating the rise of income inequality. Both of these later values were included as independent variables in all regressions.

The work of with Bennet and Vedder (2013) examined this relationship using state-level data. Their research postulates that increasing degrees of economic freedom is associated with lower degrees of income inequality. Building on the work of Bergen (1999), Scully (2002), and Ashby and Sobel (2008), they argued that the relationship between economic freedom and inequality was represented by an inverse U-curve. At low levels of economic freedom, a marginal increase will yield higher inequality. However, as economic freedom continues to rise, marginal increases will begin to lower inequality. Figure 1 below illustrates this relationship.

Figure **1**. Average Economic Freedom.

Figure **2.** The Poverty, Inequality, and Growth Triangle.

The relationship between income inequality, growth and poverty is conceptualized via Figure **2**. The interactions between poverty, income inequality, and economic growth show that when these values experience growth or decline, there is a ripple effect that reaches across all income groups (Naschold, 2003). Data regarding these values were also included in our regressions to see how they affected results.

**Data:**

The dependent variable, GINI*ij*, is a value calculated by the US Census that is specific to each county in the sample population. Where, *i*, represents the respective county in the analysis, and *j,* represents the respective census year for which the value was calculated. The analysis was composed of data collected from American Community Survey, collected by the US Census, which aggregates survey data from households in a certain geographic to confidently assign aggregate values for calculations regarding income inequality. The values are calculated using the varying income levels respective to their share of the population area. (US Census)

Data was collected at a county-wide level for Indiana, Kentucky, and Ohio, which comprises some 300 independent county locales. All data was compiled by variable and by year, with the time and state variables being controlled by factored dummy time and dummy state variables. These were controls used by Levernier and Rickman to control for regional and time effects in their pooled OLS regression. Nice link to the literature The observations were collected in 1990, 2000, and 2010, using the ACS 5-year estimate values.

Table **1.** Descriptive Statistics.

The explanatory variables, unemployment, per capita income, population, and percent of people over age 25 with a high school degree (education), were common factors used in the OLS regression suggested by Levernier and Rickman in 2013, amongst other more specific economic and demographic variables. The percent of people below the federal poverty line, Poverty*ij*, was added to see how it would affect the regression results. These variables are hypothesized to be indicators of the robustness of wealth distribution across a populace and will be tested to see how they affect GINI coefficients at a county-level.

Collecting data regarding multiple geographic areas across a common in their analysis of a state by state income inequality, Levernier and Rickman discuss how multicollinearity affected their results. They note that single period regressions have fewer degrees of freedom by nature, and that further, measures that affect both measures that labor markets and demographic variables, like education, could affect each other. Controls for these interactions will be discussed in the empirical approach section.

The models for regression were adapted from a thorough literature review which identified two major trends in this type of analysis, pooled OLS and panel regressions. The data was collected into a panel frame for the specified variables and time periods, *j*. Pooled OLS was thought to aid in the control of regional effects on results and could aid in predicting how the variables are able to affect GINI values (Levernier and Rickman).

The data for the GINI coefficient, income, education rates and poverty rates were arranged for the 3 states in a single panel data format. Quantitative data analysis such as the ADF and KPSS tests were conducted to dictate the stationarity of the data. Descriptive statistics were performed to determine the nature of the data, which is indicated in Table **1**.

The dataset was tested for several OLS assumptions to address issues such as multicollinearity, heteroskedasticity, and stationarity. It was hypothesized that the effects of time would have a correlation on results, and this was identified once graphing poverty vs. time, which is displayed in Figure **3**.

Figure **3**. Plot of Poverty vs. time.

Literature addressed for issues arising between using single period regressions by adding dummy variables for time. Multicollinearity was addressed by adding dummy variables, and by adding more defining demographic statistics regarding both economic effects such as income per sector, or income of goods producing sectors, and demographics including ethnicity and gender. These data could possibly positively affect the validity of the results and need to be investigated further. These specific regressions used in this paper used the rates of change for the entire, unspecified population.

Figure **4**. GINI values per county for 1990, 2000, and 2010.

**Empirical Approach:**

The variables income and population follow a wide distribution which ranges over several orders of magnitude thus it is crucial to take the logarithm of those variables. This transformation helps to normalize the skewness of the data. (Croissant, Milo)

The explanation behind each test was driven by the Journal of Statistical Software, Panel Data Econometrics in R: The plm Package, and Chapter 15 from Principles of Econometrics with R, Panel Data Models. The Lagrange Multiplier Test (Gourieroux, Holly and Monfort) was ran for two-way fixed effects (county and year). We reject the null hypothesis at 99.9% significance level that there are no significant two-way fixed effects with a P-value of < 2.2 e-16. Thus, there are significant two-way fixed effects. With respect to the null hypothesis of trend stationarity, the KPSS test indicates that the poverty and unemployment variables were non-stationary, while affirming that the other variables were stationary. (Kwiatoski et al, 1992) (Croissant, Milo)

Serial correlation was tested using the Breusch-Godfrey/Woolridge test with the null hypothesis being that there is no serial correlation. P-value of 0.003454 was obtained, which means we accept reject the null at 99.9% level of confidence. The Breusch Pagan test was used to check for heteroskedasticity. With a p-value of 5.609e-07, the null hypothesis of homoscedasticity has to be rejected at 99.9% confidence level. Thus, the data is heteroscedastic.

Using a pooled fixed effects OLS model was supported by Levernier and Rickman. They use this type of regression to mask the bias caused by controlling time and states??? Do you mean that a fixed effects panel model helps with bias by controlling for time and state fixed effects?. The authors also note that, the fixed effects used in their model (which was a regional analysis), could have been incomplete in their goal of controlling for state fixed effects in such a cross-section specification, which could bias results. In order to address heteroskedasticity and stationarity for the OLS regression ,I know differencing helps with non-stationarity; does it also help heteroskedasticity? the variables were differenced to an order of I(1).

With regards to the panel data analysis, a large sample size and small number of time periods fixed effects model is said to the more efficient than the first difference model. (Advanced Panel Data Methods)

The first difference (FD) model is a solution to tackle the problem of unobserved time constant variant heterogeneity of the data. When serial correlation and heteroskedasticity is present in the data the first difference (FD) model is a more efficient than the fixed effects model (Wooldridge, 2006). We decided to base the panel data analysis was established upon fixed effects and first difference models. The fixed effects also were verified as valid in the pooled OLS regressions by using ANOVA. The results from this analysis are displayed in Table **2**.

Table **2**. ANOVA.

The ANOVA results showed that the added dummy variables were significant. The added fixed effects also tried to address the possibility of omitted variable bias, and sought to mitigate the effects that regional, state, and time bias were affecting the results.

The regressions were tested via VIF testing for multicollinearity. The VIF values in the undifferenced regression had unacceptably high VIF values for two variables, log per capita income, 12.545916, and the factored dummy year variable, with a VIF value equaling 31.280846. This problem was sought to be mitigated by differencing values that literature suggesting could be causing the discrepancies, and the protocol used to correct this is listed in results. Multicollinearity in regressions using data such as single locales over specific time periods is a problem is common across literature (Levernier and Rickman), especially among education and unemployment values.

Pooled OLS Model

***ij ij ij***

Panel Model

***ij ij ij***

**Results:**

Table **3**. Pooled OLS Regressions;

Table **4.** Panel Regressions.

The results from this analysis will be described from both the pooled OLS regression and that of the panel regression perspectives.

**Pooled OLS**

Empirical estimates of the pooled indifferenced and difference OLS models are shown in table **3**. The regression estimates show factors that impact interstate state level of income inequality at a particular point in time. All variables, explicitly excluding per capita income, unemployment and the state dummy for Ohio, are significant with p-values < 0.01. The estimates along with the value of the R^2 show that the investigated variables on average explain a considerable share of interstate differences in state income inequality between the years 1990-2010. The value of the time dummies become more negative over time suggesting that income inequality decreased with as time moved on. The magnitude for the coefficient of the state dummy for Kentucky is the highest, indicating that the marginal effect on the average levels of the GINI coefficient is the highest, which supports the empirical findings by the Center of Budget and Policy Priorities indicating that Kentucky is among the states with the highest level of income inequality. I will accept this as a comparison of the parameter estimates, but I want you to correct this for the final project. It turns out you can’t compare the estimates of dummy variables to the estimates of continuous variables. Parameter estimates for dummy variables are interpreted relative to the omitted category, like the Kentucky estimate is how much different Kentucky’s GINI is than Indiana’s, all else constant. For the final project, you can compare Kentucky’s estimate to Ohio’s, or calculate what the largest elasticity is among the continuous variables

The OLS regression that was deemed to be the best fit model was that for which the nonstationary variables, poverty, and unemployment were differenced to the first order, I(1). Both models were fixed effects models after ANOVA showed its significance. As discussed previously, KPSS testing showed these variables to be the most affected by stationarity. Other values were trimmed to 899 observations to avoid variable length errors. Wow, you thought of everything! Although both the R^2 and AIC/BIC statistics for this regression were lower than that of the non-differenced model, VIF testing confirmed that multicollinearity was significantly decreased by differencing the nonstationary variables. This led to the conclusion that the differenced model was overall more accurate.

After variable treatment, the most significant effects of changes in GINI were population, education, and poverty, while also both regional and time dummies maintaining a presence through marginal effects. Income per capita and unemployment became less significant once treated. The role that these two later variables play in determining their effect on GINI values although insignificant in the OLS model, seemed to play a role in determining GINI from a panel data aspect.

The both regressions showed negative elasticities in the educational category, which was expected. Higher education rates were expected to decrease per unit GINI, and under inspection this relationship has validity. As noted in previous research, as this value increases there is a negative effect on income inequality (Sale, 1974).

Income per capita was found to be insignificant in the final OLS regression. This could be because income per capita is not a robust enough measure of the real share of income between the population of a particular county. Research into this topic typically calculates their own statistics for analyzing real share of income in a geographic region, (Bennet, Vedder, 2013). Considering this, further investigation into the GINI coefficients when considering a dataset like ours would need to consider the pros and cons for determining which value to use for a true median income value.

The first differenced OLS model shows that all the variables are significant other than the unemployment rate, log income and the state dummy variable for Ohio. Poverty and income inequality share a negative relationship. An increase in one unit of the poverty rate would cause the Gini coefficient to decrease by 0.07148 units. This negative relationship is demonstrated by the poverty-growth-income triangle. In an area with high income inequality, economic growth leads to a reduction in poverty rates and exacerbates the level of income inequality. (Bourguignon, 2004)

**Panel**

The dependent variable in all the fixed effects regressions is the GINI coefficient. We started with all the variables with the fixed effects model. It was determined that all the variables other than the education variable were statistically significant. The coefficient of the independent variable indicates how much the GINI coefficient changes over time on average per county when the independent variable increases by one unit. Considering the logged independent variables an increase in 1-unit GINI to of the independent log variable leads to a percentage change in income inequality determined by 100 times the beta coefficient on average over time. An increase in one unit of log income would lead to an increase in 63% of the GINI. -3 unless I’m mistaken, doesn’t 0.63 parameter estimate mean increasing income by one *percent* is associated with a 0.63/100 = *0.0063* *percentage point* increase in GINI coefficient *all else constant*? Shame to have to take off any points for such a thoroughly done study

An interesting finding of this analysis is that income inequality is positively correlated with the county level incomes. Our results align with the findings of Partridge (1997) who used a panel of 48 US states to show that state level incomes were positively correlated with income inequality. His hypothesis was the first one to contest the idea that income inequality was harmful for economic growth. Income inequality increases in times of high economic growth.

The first difference model indicates that the log (income), education, poverty and the state dummies as being significant. Compared to the fixed effects estimation, all the variables have a higher magnitude compared to the fixed effects model after removing the unobserved time heterogeneity. The education rate has a negative coefficient indicating that a unit increase in the education rate you should say what the units are: percentage points, so increasing educational attainment by one percentage point would decrease the GINI coefficient by 0.0729/100 = 0.0007 units percentage points on average all else constant. As economic growth is more robust the gap between the rich and poor is increasing. This leads to higher education being accessible to families with higher incomes while students from low income households are less likely to go to college and to graduate. (Hill, 2015)

**Conclusion:**

The paper examined the relationship between the commonly used measure of county income inequality, the Gini coefficient and numerous county specific economic variables such as population, income and unemployment along with time and state wise dummies. The results of the previous studies were updated as it specifically examined the county wise income inequality. The OLS results show that the factors that contribute to changes in income inequality in our neighborhood states were education and poverty. Furthermore, it also shows that a county is more likely to have higher income inequality if it has lower education rates, marginally lower levels of poverty and is situated in Kentucky, this is visualized for our population set in Figure **4**. The panel results show income inequality within a county over time is attributed to the log per capita income levels, poverty rates, education rates and the state that it is located in.

The data looks to be clustered into years, so we will continue to investigate how this could be addressed by running different type of regressions, such as spatial dependence model to account for the clustering.

This has several policy implications. Economic growth enhancing reforms should be implemented so that lower income households have access to higher levels of disposable income. Higher levels of disposable income would give households access to better education, healthcare and lower rates of poverty. This would narrow the income inequality in disposable income. (OECD, 2015)

For the final project, fix the errors above, then do the following for your fixed effects panel model. YOu will probably want to do these in separate regressions.

1. Add an interaction term between education and poverty. Interpret the relationship between education and GINI in light of the interaction term
2. Run a quantile regression. Compare the relationship between education and GINI at the 25th and 75th percentiles of income inequality. You can do this in a one-year cross section if you want, but there are panel models available
3. See if your model follows a finite mixture distribution and what you can learn by running it as a finite mixture model. You can do this in a one-year cross section, too.

**Works Cited:**

*Advanced Panel Data Methods*[Scholarly project]. (n.d.). In *Montana State University*. Retrieved from http://www.montana.edu/cstoddard/562/Chapter14studentnotes.pdf

Bennett, D. L., & Vedder, R. K. (2012). A Dynamic Analysis of Economic Freedom and Income Inequality in the 50 U.S. States: Empirical Evidence of a Parabolic Relationship. *SSRN Electronic Journal*. doi:10.2139/ssrn.2134650

Bourguignon, F. (2004). *The Poverty-growth-inequality triangle*(Working paper No. 125). New Delhi: Indian Council for Research on International Economic Relations (ICRIER).

Croissant, Y., & Millo, G. (n.d.). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*.

Frank, M. W. (2009). Inequality And Growth In The United States: Evidence From A New State-Level Panel Of Income Inequality Measures. *Economic Inquiry,47*(1), 55-68. doi:10.1111/j.1465-7295.2008.00122.x

Galili, T. (2013, May 27). Log Transformations for Skewed and Wide Distributions. Retrieved from https://www.r-statistics.com/2013/05/log-transformations-for-skewed-and-wide-distributions-from-practical-data-science-with-r/

Glassman, B. (2017). Income Inequality among Regions and Metropolitan Statistical Areas: 2005 to 2015. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2017/demo/SEHSD-WP2017-41.pdf

Grundy, P. (2017). Pulling Apart. *Color and Character*. doi:10.5149/northcarolina/9781469636078.003.0006

Hill, C. B. (2015). Income Inequality and Higher Education. Retrieved from https://www.acenet.edu/the-presidency/columns-and-features/Pages/Income-Inequality-and-Higher-Education.aspx

Iii, T. S. (1974). Interstate Analysis of the Size Distribution of Family Income, 1950-1970. *Southern Economic Journal,40*(3), 434. doi:10.2307/1056017

*INCOME INEQUALITY HAS GROWN IN KENTUCKY*(Rep.). (n.d.). Retrieved https://www.cbpp.org/sites/default/files/atoms/files/Kentucky.pdf

Kuznets, S. (1955). Economic Growth and Income Inequality. *The American Economic Review,45*(1). Retrieved from https://www.jstor.org/stable/pdf/1811581.pdf?refreqid=excelsior:5a1b7a97a79b79320f9062cb058a5a8e&seq=1#page\_scan\_tab\_contents.

McDonough, P., PhD, Duncan, G. J., PhD, Williams, D., PhD, & House, J., PhD. (1997). Income Dynamics and Adult Mortality in the United States, 1972 through 1989. *American Journal of Public Health,87*(9). Retrieved from https://ajph.aphapublications.org/doi/pdf/10.2105/AJPH.87.9.1476.

Naschold, F. (n.d.). *Why Inequality Matters for Poverty*(Issue brief).

Saez, E. (2016). *Striking it Richer: The Evolution of Top Incomes in the United States (Updated with 2015 preliminary estimates)*(Tech.). University of California, Berkeley.

US Census Bureau. (2016, January 25). Income Inequality. Retrieved from https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html

Variation in U.S. State Income Inequality: 1960-1990. (n.d.). Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/016001769501800305

Variation in U.S. State Income Inequality: 1960-1990. (n.d.). Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/016001769501800305?casa\_token=408gLj3fP-8AAAAA:j0EGJguJj\_u2pd3HGQQhDDDRsYuqqf1JhnblR4eylm4\_9Ifwqk-OtZARhD5jv-b-nti2wmYd5w

Wooldridge. (2006). *Introductory Econometrics (3rd ed)*. Cincinnati, OH: South-Western, Cengage Learning.

**Data and Figures:**

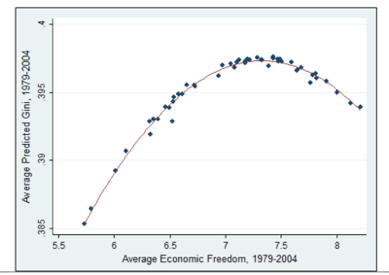
**

Figure **1**.

GINI coefficients as a function of Average Economic Freedom (1979-2004).

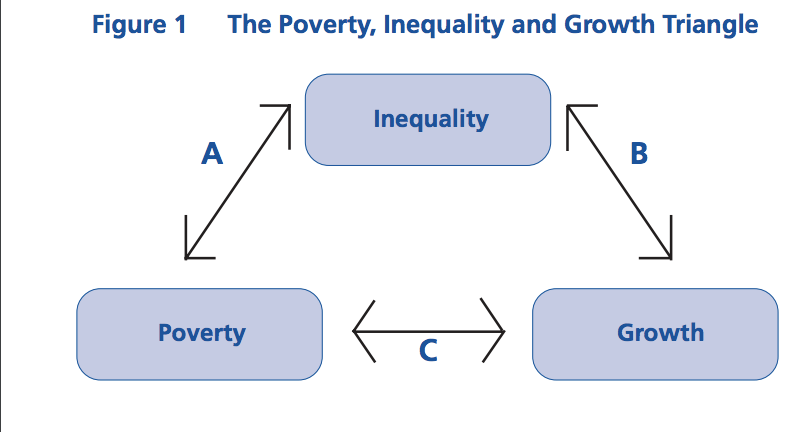


Figure **2**, The Poverty, Inequality, and Growth Triangle.

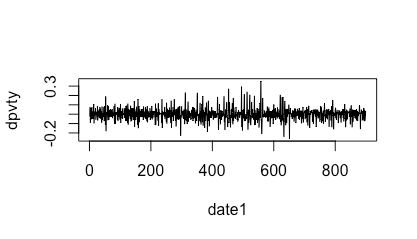
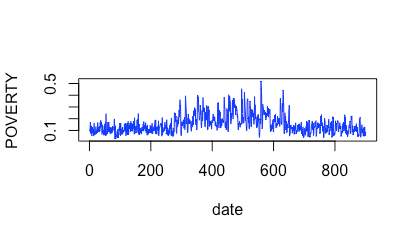
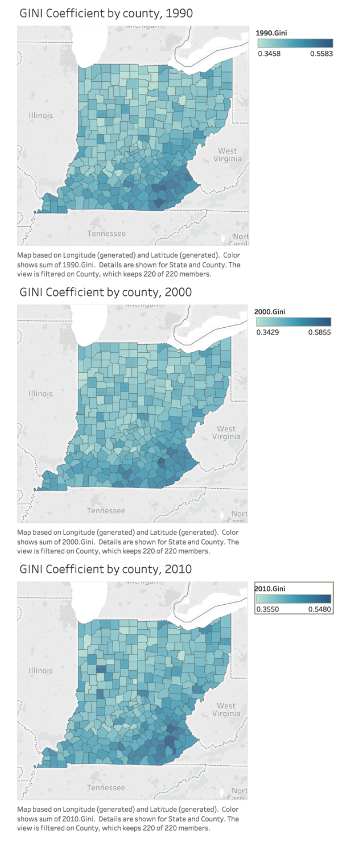


Figure **3**. Plot of Poverty vs. time

A plot of poverty versus time, where observations (date) 1-300 represent values in 1990, 301-600 represent values in 2000, and 601-900 represent values in 2010. There are three visible cluster which represent nonstationarity and spatial dependence. The second chart (date1) represents the poverty values post-differencing.

 Figure **4.** GINI by county for 1990, 2000, and 2010.

Kentucky has a disproportionate number of

counties with high GINI compared to that of Indiana

and Ohio

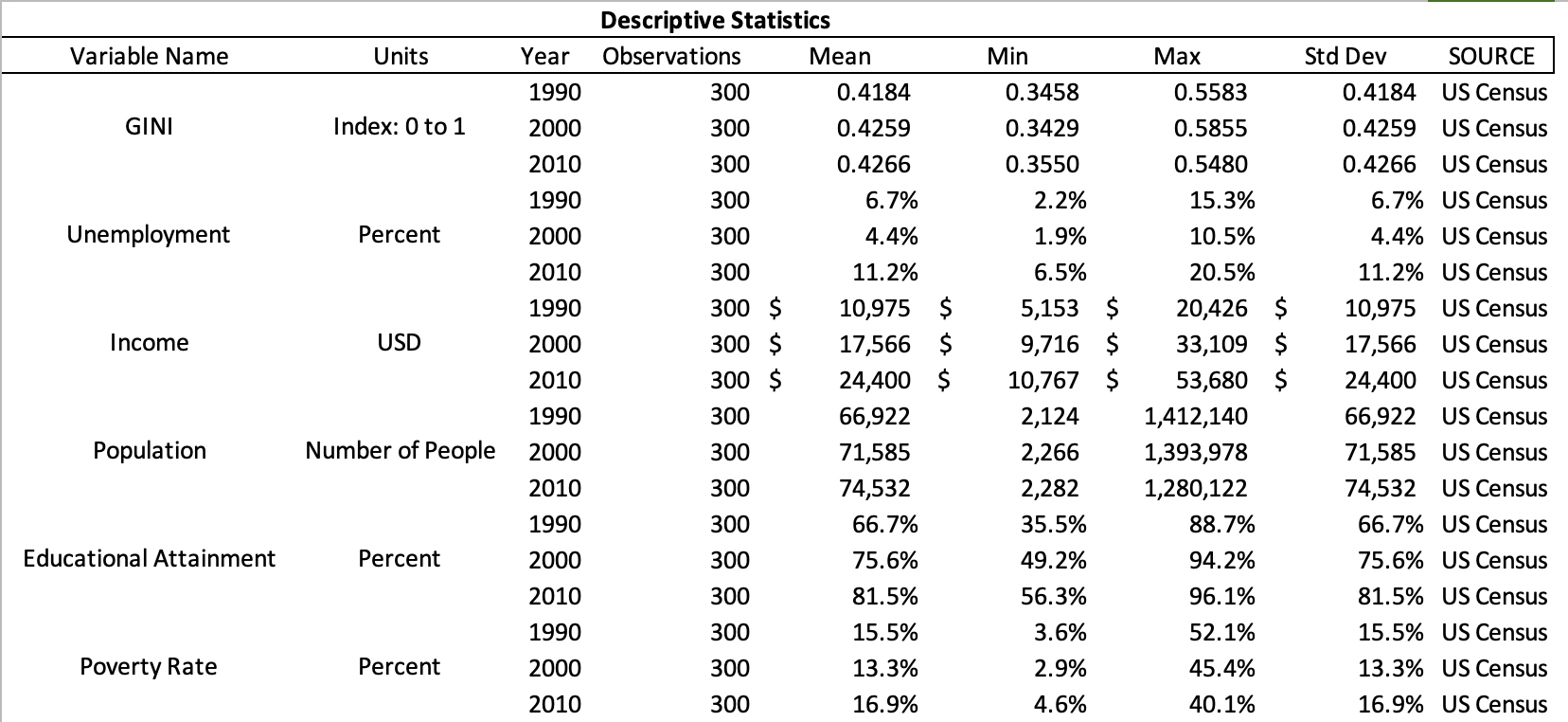


Table **1.**

This table contains summary statistics for the dataset which was used in the regressions. The data is was arranged vertically, as opposed to its horizontal representation above, and was used with dummy variables to distinguish between state and time. The 300 observations above include summary values for each county, collectively. The data was obtained by using the US Census’ Factfinder tool.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | RSS | Df Sum of Sq | F | p-value |
| 1 (without FE) | 0.47033 |  |  |  |
| 2 (with FE) | 0.42184 | 0.048485 | 25.573 | <2.2e-16\*\*\* |

P-value significance codes: “ ` ” p<.1 “ \* ” p < .05 “ \*\* ” p < .01 “ \*\*\* ” p < .001

Table **2**. ANOVA

The above table was the results of an ANOVA analysis used to test the significance of the added fixed effects (FE), time period (1990, 2000, and 2010) and state (Indiana, Kentucky, and Ohio). The results show that there is a significant correlation between the the model with FE to that of the control model which excludes them.

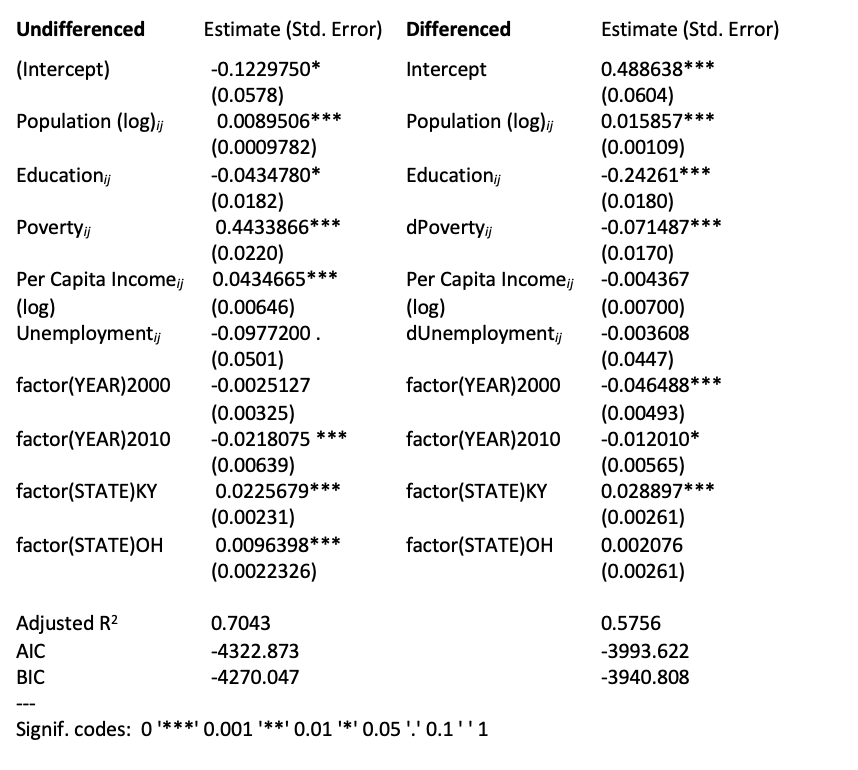
****

Table **3.** OLS Regressions (differenced and undifferenced models)

The above table shows regression results for the two OLS restricted and unrestricted models that were analyzed during the investigation. Differenced variables are indicated with a d prefix.

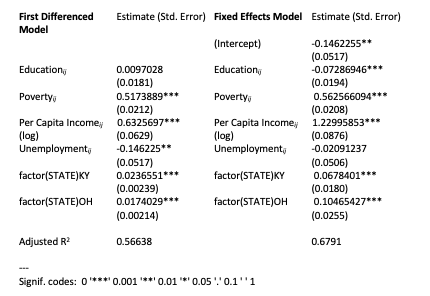


Table **4**. PLM Regressions

The above table shows regression results for the fixed effects and first difference estimations. The two-way fixed effects capture marginal effects of county and time.