NCED: Part 3

Christian Okoth

December 5, 2019

## Summary

North Carolina is a large state with urban and rural counties. Both are essential to NC’s economic development. This study aims to uncover which counties are at the higher and lower ends of economic developoment by studying income and population effects in each county, regressing a number of development indicators on median family income. I am using family income as a proxy for how developed a county is, but it is important to note I also could have used wage. Below is a key of my variables.

Key:

Variable name [#] Description of variable

Area [1] "Area Name"

Medi\_owner [2] "Median Value of Owner Occupied Units"

Outsiders [3] "Persons Living Outside This County Five Years Ago"

pop [4] "Population Estimate (BEA per Capita Denominator)"

avg\_employ [5] "Average Annual Employment by Place of Work"

avg\_wage [6] "Average Annual Wage per Worker"

employ\_resid [7] "Employment by Place of Residence"

fam\_income [8] "Estimated Median Family Income(HUD)"

college [9] "College Graduates Age 25 Up"

high\_sch [10] "High School Graduates Age 25 Up"

sat [11] "SAT Grand Total Average Score"

white\_maj [12] "Majority White"

part [13] "Part of the State"

log\_pop [14] "Logarithm of Population"

int\_own\_wage [15] "Median Value of Units Annual Wage Interaction"

## Full Regression

Model 3: OLS unrestricted, using observations 1-100

Dependent variable: fam\_income

Heteroskedasticity-robust standard errors, variant HC1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| const | 4847.78 | 12880.5 | 0.3764 | 0.7076 |  |
| medi\_owner | 0.115343 | 0.0194564 | 5.928 | <0.0001 | \*\*\* |
| avg\_employ | −0.160437 | 0.0500124 | −3.208 | 0.0019 | \*\*\* |
| employ\_resid | 0.509808 | 0.183596 | 2.777 | 0.0067 | \*\*\* |
| college | −0.419868 | 0.155349 | −2.703 | 0.0083 | \*\*\* |
| part | 4863.04 | 1641.88 | 2.962 | 0.0039 | \*\*\* |
| outsiders | −0.492477 | 0.314998 | −1.563 | 0.1215 |  |
| pop | 0.0384411 | 0.113381 | 0.3390 | 0.7354 |  |
| avg\_wage | 0.469892 | 0.166055 | 2.830 | 0.0058 | \*\*\* |
| high\_sch | −0.455035 | 0.271581 | −1.676 | 0.0974 | \* |
| sat | 4.82509 | 8.75746 | 0.5510 | 0.5831 |  |
| white\_maj | 2256.15 | 2074.33 | 1.088 | 0.2797 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 53993.00 |  | S.D. dependent var | 9201.661 |
| Sum squared resid | 2.33e+09 |  | S.E. of regression | 5142.316 |
| R-squared | 0.722392 |  | Adjusted R-squared | 0.687691 |
| F(11, 88) | 45.26854 |  | P-value(F) | 1.80e-31 |
| Log-likelihood | −990.0281 |  | Akaike criterion | 2004.056 |
| Schwarz criterion | 2035.318 |  | Hannan-Quinn | 2016.708 |

Regression on family income: the mean for the sample of family income is $59,993.00 annually, and the standard deviation is $9,201.66. Of my regressors, 7 were statistically significant, and 6 were significant at the alpha = .01 level. Median value of owner occupied units was positive and statistically significant at the alpha = .01 level, as were employment by residence, part of the state, and average wage by employment. Average employment by place of work was negative and statistically significant at the alpha = .01 level, as were college graduates. The only variable to be significant at the alpha = .1 level was high school graduates, which entered the regression as negative. The results tell me that employment statistics, the value of an average home, and education are all significant predictors of family income. The standard error for the regression was $5142.32, and reports a p-value > .0000 with an F(11,88) distribution. Multicollineatiry may have played a part in my results as well, as many of my variables were correlated with my dependent variable, as well as my independent variable.

## F-Test for Joint Hypothesis

Model 4: OLS restricted, using observations 1-100

Dependent variable: fam\_income

Heteroskedasticity-robust standard errors, variant HC1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| const | 16613.0 | 5947.67 | 2.793 | 0.0063 | \*\*\* |
| medi\_owner | 0.128122 | 0.0117911 | 10.87 | <0.0001 | \*\*\* |
| avg\_employ | −0.208827 | 0.0520809 | −4.010 | 0.0001 | \*\*\* |
| employ\_resid | 0.561462 | 0.161265 | 3.482 | 0.0008 | \*\*\* |
| college | −0.410620 | 0.143674 | −2.858 | 0.0053 | \*\*\* |
| part | 2750.08 | 1104.43 | 2.490 | 0.0146 | \*\* |
| avg\_wage | 0.393614 | 0.172344 | 2.284 | 0.0247 | \*\* |
| high\_sch | −0.422651 | 0.179294 | −2.357 | 0.0205 | \*\* |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 53993.00 |  | S.D. dependent var | 9201.661 |
| Sum squared resid | 2.53e+09 |  | S.E. of regression | 5238.878 |
| R-squared | 0.698771 |  | Adjusted R-squared | 0.675851 |
| F(7, 92) | 46.22289 |  | P-value(F) | 1.94e-27 |
| Log-likelihood | −994.1110 |  | Akaike criterion | 2004.222 |
| Schwarz criterion | 2025.063 |  | Hannan-Quinn | 2012.657 |

H\_{0}: beta3 =beta4 = beta5 =beta6 = 0 H\_{A}: at least one beta*i* = 0

where restricted is model 4, unsrestricted is model 3 F-test:

F = {{SSE\_{R}-SSE\_{U}}/J}/{SSE\_{U}/{T-K}}

{{2.53e+09-2.33e+09}/4}/{2.33e+09/{88}} = 1.888412

Fail to reject H\_{0}. There is insufficient evidence at the alpha = .1 level to suggest that beta3 =beta4 = beta5 =beta6 do not equal 0. Essentially, there is no reason to conclude that these variables add signifcantly to the explanatory power of the model. This result makes sense in context as the adjusted r^{2} are different from each other by 0.01184, or 1.18%. The SSR are not substantially different either.

My initial regression and the results of my F-Test found that none of the variables I thought would be significant were particularly significant, which does not line up with my initial hypothesis that population would be correlated with income. Surprisingly, I discovered that the racial makeup and population size of a county were not significant variables, nor was the variable measuring how many people had moved into the county in the last five years. I expected these to have a stronger effect in the regression, because these indicators measure some of the basic demographics of the workforce. While I expected the more obvious economic indicators to be significant, I was surprised by the insignificance of the labor force variables that my model predicted.

The correlations I found in Part 2 were not supported by the regression output, likely because of omitted variable bias, or other variables that I did not include in my regression that may have biased the estimatiors in my regression. The model was most susceptible to bias in my estimators for income, as they were measures of average wage and household income, neither of which can measure income accurately by itself. I also included some terms which were highly correlated with one another, especially the SAT score variable, which was highly correlated with all of my variables (it was not statistically significant). According to my correlation matrix, all of my variable could have been significant predictors of family income.

## Measures of Fit & Assumptions

In terms of the two measures of fit, how well does the model ﬁt the data? Discuss whether your model is appropriate given how well it does or does not ﬁt the data. In terms of r^{2} value of .728104, my regression explains about 73% of the variation in y, family income. The standard error for the regression is 5147.973, meaning that on average, the deviation from the regression line was $5,147.97. These measures say that the regression overall can explain a significant amount of variation in the data but is not necessarily the best fit for the data. A lower SER would be preferred.

To maintain statistical validity, it is important to assess the required conditions for regression.

Assumptions:

1. E[u\_{i}|X\_{1i}…X\_{ki}] = 0 is verified by the residuals plot. There is no clear pattern to the data, verifying random assignment.

A screenshot of a social media post

Description automatically generated

A close up of a map

Description automatically generated

1. (X\_{1i}, X\_{2i},…,X\_{ki},Y\_{i}) i=1,…,n are i.i.d cannot be verified because the sampling design is not explicity random. If the sampling earlier in the data pipeline is good, then this assumption would be satisfied.
2. Normal enough and no large outliers is verified by the histogram and QQ-plot indicating strong normality of the dependent variable.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

1. No perfect multicollinearity was verified in earlier paper.

## Nonlinearity

I chose to use two nonlinear terms, the logarithm of population and an interaction term to measure the interaction between the average cost of a home and the average wage per worker. I based the log off of the scatterplot of population against family income because it seemed to follow an exponential growth pattern.

A screenshot of a cell phone

Description automatically generated

I chose the interaction term based upon evidence from my correlation matrix, as average cost of a home and average wage per worker had the smallest p-value of any correlation (p-value = 3.265641e-04) that was not a correlation with the dependent variable.

A graph plotting medi_owner, avg_wage, and the interaction term against fam_inc



Controlling for nonlinearity is essential when your variables model nonlinear processes and help to standardize your residuals. Rather than having low residuals for low values and high residuals for high values due to incorrect fitting, the data are adjusted to be modeled better alongside linear variables.

Model 1: OLS including nonlinear terms, using observations 1-100

Dependent variable: fam\_income

Heteroskedasticity-robust standard errors, variant HC1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *Coefficient* | *Std. Error* | *t-ratio* | *p-value* |  |
| const | −10400.2 | 29092.4 | −0.3575 | 0.7216 |  |
| medi\_owner | 0.219269 | 0.107887 | 2.032 | 0.0452 | \*\* |
| outsiders | −0.317025 | 0.339230 | −0.9345 | 0.3526 |  |
| pop | −0.00638567 | 0.118927 | −0.05369 | 0.9573 |  |
| avg\_employ | −0.191406 | 0.0605513 | −3.161 | 0.0022 | \*\*\* |
| avg\_wage | 1.04641 | 0.643786 | 1.625 | 0.1077 |  |
| employ\_resid | 0.544148 | 0.190685 | 2.854 | 0.0054 | \*\*\* |
| college | −0.336278 | 0.170212 | −1.976 | 0.0514 | \* |
| high\_sch | −0.334612 | 0.290797 | −1.151 | 0.2531 |  |
| sat | 7.54069 | 9.17277 | 0.8221 | 0.4133 |  |
| white\_maj | 2147.07 | 2069.57 | 1.037 | 0.3024 |  |
| part | 4950.25 | 1645.43 | 3.008 | 0.0034 | \*\*\* |
| log\_pop | −729.275 | 1500.58 | −0.4860 | 0.6282 |  |
| int\_own\_wage | −3.2310e-06 | 3.4827e-06 | −0.9277 | 0.3561 |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mean dependent var | 53993.00 |  | S.D. dependent var | 9201.661 |
| Sum squared resid | 2.28e+09 |  | S.E. of regression | 5147.973 |
| R-squared | 0.728104 |  | Adjusted R-squared | 0.687003 |
| F(13, 86) | 3.840011 |  | P-value(F) | 0.000072 |
| Log-likelihood | −988.9885 |  | Akaike criterion | 2005.977 |
| Schwarz criterion | 2042.449 |  | Hannan-Quinn | 2020.738 |

Having rerun my regression, my original model has an adjusted r^{2} value of .6877 and a SER of $5142.32. Comapatively, my nonlinear adjusted model has an adjusted r-square of .687, meaning that the value of the added nonlinear terms did not increase the explanatory power of the model relative to the original model. In other words, the benefit of the additional terms was completely offset by the adjustment in degrees of freedom related to the number of regressors. The added nonlinear terms failed to change the explanatory power, likely because both terms individually were not statistically significant, or those processes were not actually modeled after nonlinear relationships. Further, the SER of my nonlinear adjusted model was 5147.97, representing an increase in the variability of the model, rather than a decrease, indicating that the model is a worse fit for the model relative to the number of regressors. My original model is the better model, because it more efficiently predicts the variation in the data.

## Summary

My analysis found that there are multiple development indicators that can act as a predictor for economic development. While family income may have been a flawed indicator, the regression did show statistically significant relationships between family income and the median value of owner occupied units, average rate of employment, employment by residence, college graduates, and the part of the state.

I would address the governor, noting that income has a strong relationship with a number of other economic indicators, and that while it is essential to focus on developing each and every one, the variables have strong relationships with one another, so it is best to act in multiple sectors at once. I would suggest focusing on education and employment. My regression found that the most significant variable effects had to do with whether people were employed and how well they were being compensated, and how well educated they are, so whether or not a significant number of people have graduated high school and college. The wisest investments based upon my model would be to invest in attracting college graduates to less developed areas, and making sure people are employed.