NCED: Part 3

Christian Okoth

December 5, 2019

## Header 1: Review

Give me a brief introduction about the nature of your topic—remember that I do not have the earlier parts of your project. 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.

1. Run a regression of your dependent variable on all of your independent variables. Interpret the regression by providing a verbal explanation of the results. Make precise statements about what your results tell you, being sure to note which coefficients are statistically significant and which are not.

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 = .01 level. Median value of owner occupied units was positive and statistically significant at the = .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 = .01 level, as were college graduates. The only variable to be significant at the = .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.

1. Run F-tests on groups of insignificant variables to see if the null hypothesis that all the coefficients in the group are jointly equal to zero.

H\_{0}: *{3} =* {4} = *{11} =* {12} = 0 H\_{A}: at least one \_{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 = .1 level to suggest that *{3},* {4}, *{11} &* {12} 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.

1. Do the regression results support the predictions you discussed in the earlier portions of the project? Explain. What did you ﬁnd that was surprising? The regression 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 of a county were not significant variables, nor was the variable measuring how many people had moved to 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.

Do the regression results support or refute the correlations you found in PA#2 question 1.c? Explain why or why not. Relate your discussion to our class discussion of omitted variable bias. 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.

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.

## Header 2: Assumptions and Nonlinearity

1. Check the required conditions for regressions by performing a residual analysis. You can save the residuals and create a histogram to check for normality and outliers by examining the residuals from your regression. Checking the summary statistics for your residuals may be helpful as well.

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.

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. 4.No perfect multicollinearity was verifed in earlier paper.

Create a nonlinear function of a single independent variable (Section 8.2). A popular choice is the square of one of your regressors (but it doesn’t have to be). Pick which X to use by looking at the scatterplots of your dependent variable against each of your continuous variables to ﬁnd a relationship that appears nonlinear. You may also choose a log transform of one or more variables or use an interaction term. 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. I chose the interaction term based upon 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 with the dependent variable.

Discuss why controlling for this type of nonlinearity is important and might improve your results. Rerun your regression model with the nonlinear term(s) included. Discuss any changes in the results you observe. Which model is preferred?

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. 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

1. Summarize your project. What are the policy implications of your ﬁndings? Your discussion should be in terms that would be helpful to a policy maker relevant to your research question.

I would address the govenor, 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.