Elliott Evans, Raymond Luu
BIOSTATS 696
11/28/2016

Spatial Analysis of Michigan High School Dropout Rates

Introduction

Completion of high school education has been a well documented predictor of life success. Naturally, those whom leave high school early usually do not qualify for college education, and this limits the amount of the fields that careers that are available to them. This is reflected upon statistics such as unemployment rate, in which it could be seen that on average across the United States, joblessness among those who lack a high school degree is an enormous 12 percent, as opposed to the 8.1 percent national average. The rate among those whom graduated college shrinks to down to 4.1 percent [1]. Additionally, the salaries of this demographic tend to be drastically limited. For those over the age of 25, the average weekly income for someone who has not completed high school is only 493 dollars. Completion of high school pulls this average to 678 and a college degree skyrockets it to 1137 [2]. Thus it is imperative to ensure that youth succeed in completing their high school education. However, there are great variations across the country in terms of high school graduation rate. This rate ranges from a fantastic 90.8 percent rate in Iowa to a troubling 68.5 percent rate in the District of Columbia [3]. Michigan ranks in the lower half of this spread at 79.8 percent which acts as our motivation for this study. Graduation rates spatially vary across the nation on a state level and we suspect that this rings true on a smaller scale across Michigan on a county level.

Data

We obtained our information from the school data publically available on Michigan's official website and conduct several analyses with the intent of constructing a model from which we could derive meaningful trends. In particular, our dataset covers the 2014 to 2015 school year. The observations are at the county level and the variables we are concerned with are as follows: number graduating, cohort count, percent gender, percent economically disadvantaged, and percent race. Economically disadvantaged here is defined as a student who has been reported as eligible for supplemental nutrition in any certified collection in the MSDS in the current school year by any reporting entity or directly certified or reported as homeless or migrant.

As a disclaimer, these are a few public database discrepencies present in our data that prevent it from actually reflecting the 79.8 percent graduation rate mentioned in the introduction. Specifically, the aggregated county-level cohort count underestimates the true state level cohort count that Mighigan records, which leads to an inflated 84 percent graduation rate in the data. In the case of Keweenaw County, we imputed with means. This error does not appear to be reconcilable in a reasonable logistical manner but for the purposes of our study, this would suffice as far as finding spatial correlation is concerned.

Analyses and Results

We begin our analysis a simple choropleth map of graduation rates by Michigan count as a preliminary observation of any spatial variability

ADD 1ST FIGURE

Immediately, we could see that there indeed seems to be some spatial patterns in the country. The lowest graduation rates are bunched together near the middle of the figure while the highest ones seem to gather together near the top of Michigan. The quantify this observation, we conduct Moran's I test under randomization and yield a Moran I statistic of 0.123242912 and a corresponding p-value of 0.02883. This is statistically significant and thus the positive Moran I indicates strong spatial correlation.

Springboarding from this, we decide to study possible large scale trends in demographic data to see if these apparent spatial correlations could be explained by other factors. We fit a model using all the previously mentioned characteristics as covariates and then employsed stepwise selection with AIC as our criteria for strong predictive value. This process left us with percent black and percent economic disavantaged as our best covariates. We then refit the linear model after centering the covariates so we can easily interpret the intercept as mean graduation rate for the average proportions of blacks and economic disadvantage students in a county. Taking Y as the mean graduation rate, X_1 as the percent economically disadvantaged, and X_2 as the percent black, the model that result is as follows:

$$Y = 84.4337 - 0.1450X_1 - 0.2505X_2$$

The direction of the coefficients suggests that greater proportions of blacks and economically disadvantaged students lead to lower graduation rates. We next create figures to see if there is a spatial pattern to these covariates.

ADD 2ND FIGURE

ADD 3RD FIGURE

The visualizations make it apparent that neighboring counties for those with large black or economically disadvantaged populations tend to also respectively have large black or economically disadvantaged students. However, when we conduct Moran's I test on the residuals of the model, we find a still positive and significant statistic of 0.118629523 and p-value 0.03336, suggesting that there are still spatial variation that our covariates alone cannot explain.

To follow up on these figures, we next fit the Bayesian hierarchical spatial linear model with an improper CAR prior. We used a burn in size of 10000 on a sample of 60000 and yield the following as our summary results and trace plots.

Covariate	Median	2.5	97.5
Intercept	83.94	82.69	85.19
Percent Economic Disadvantaged	-0.14	-0.23	-0.04
Percent Black	-0.23	-0.37	-0.87
nu2	28.08	20.91	38.85
tau2	0.014	0.0027	0.495
rho	1.00	1.00	1.00

ADD TRACE PLOT FIGURES

Figures of the posterior median and standard deviation of the spatial random effects could also be found below. From these figures we see clearly that the highest posterior medians are found at the top of Michigan and the least of which are bunched together near the middle, where our preliminary graduation plot exhibits the lowest graduation rates. We also see that the greatest standard deviations are at the top of Michigan as well and the least around the lower center of Michigan. However, when we construct our credible interval for each of the 83 counties, we find that all intervals contain 0 which indicate that there are no significantly positive or negative spatial random effects in play here.

ADD POSTERIOR FIGURES

Naturally, the next step to take here would be the Bayesian spatial linear model with a proper CAR prior which yields the following as our summary results.

Coefficients	Estimate	Standard Error	P-Value
Intercept	84.47	0.775	< 2.2e-16
Percent Economic Disadvantaged	-0.114	0.0526	0.0289
Percent Black	-0.275	-0.0717	-0.0001

The values of this procedure are very similar to that with the improper CAR prior. The covariates are all significant and this model suggests a slightly stronger effect on black percent and a weaker effect on economic disadvantage.

And lastly for this family of modeling, we fit with the SAR model with nearly identical summary results.

Coefficients	Estimate	Standard Error	P-Value
Intercept	84.288	0.745	< 2.2e-16
Percent Economic Disadvantaged	-0.121	0.05155	0.0184
Percent Black	-0.275	-0.0715	-0.0001

Moving onto another model approach, we fit the spatial model to Poisson data with both spatial and non-spatial random effects. It's a Poisson model for observed counts with mean equal to expected counts times relative risk. We use S.CARbym with a burn in of 30000 and a sample of 170000 to relate the log mean of the Poisson distribution to the covariates. As a result, we use the log of expected counts as an offset when modeling the how log of relative risk varies with the covariates. Here, since we assume that the log relative risk depends on the percentage of population involved in Agriculture, Fisheries and Forestries, we have as our summary results and trace plots the following.

Covariate	Median	2.5	97.5
Intercept	-0.1566	-0.1691	-0.144
Percent Economic Disadvantaged	-0.0016	-0.0033	0.0002
Percent Black	-0.0016	-0.0033	0.0002
tau2	0.023	0.0012	0.0049
sigma2	0.0013	0.0008	0.0023

We next visualized the estimated propensity to graduate through another cloropeth. The propensities are largely consistent with our preliminary graduation rate by county cloropeth: highest near the top of Michigan and lowest near the middle.

ADD PROPENSITY FIGURE

And another posterior median of spatial random effects figure was also generated. Compared with the previous one, this is less uniform across large patches of counties. The same general trend comes across though: positive effects at the top and negative effects near the middle, though the medians in this figure deviate from 0 less, indicating weaker spatial random effects. However, we see that this time we actually do have significant spatial random effects in two counties, both in the middle portion of Michigan.

ADD NEW POSTERIOR MEDIAN FIGURES

Conclusion

References

[1] Bureau of Labor Statistics (2012). Employment Situation Summary. Retrieved from [http://www.bls.gov/news.release/empsit.nr0.htm].

[2] Bureau of Labor Statistics (2015). Earnings and unemployment rates by educational attainment. Retrieved from [http://www.bls.gov/emp/ep_chart_001.htm].

[3] U.S. Department of Education, EDFacts/Consolidated State Performance Reports (2015)