The Effect of Racial Diversity on High School Graduation Rates

Sanjay Hariharan

Department of Statistics, Duke University

Objectives

Assess what economic, social, and racial factors lead to high and low graduation outcomes across school districts in the United States:

- What factors account for the high variation in graduation rates across different counties?
- How strong is the relationship between racial diversity in a classroom and its effect on education outcomes?
- Assess migration rates in various states or cities, predict effect that diversity would have on graduation rates.
- How to best perform variable selection on a large number of covariates?

Introduction

We are interested in exploring the relationship between racial diversity in a classroom and its effect on graduation rates.

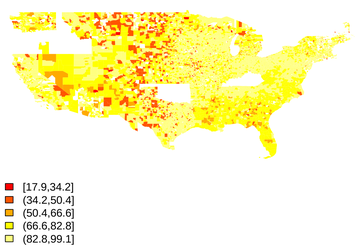
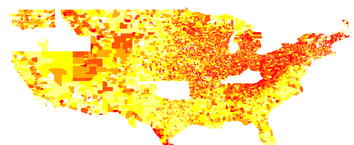


Figure 1: Graduation Rate Heatmap



Methods

We performed analysis in the following manner:

- Multivariate Data Imputation using Mice in R
- Assessed Model Fit using Variety of Different Models
- Inferred Relationship between Diversity Statistic and Graduation Rate

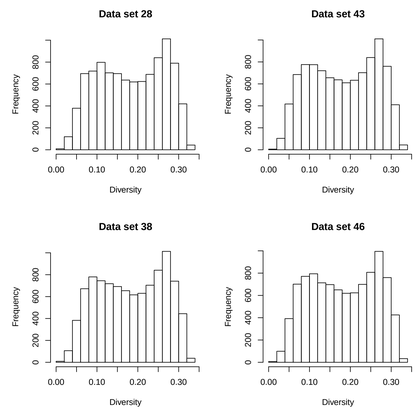


Figure 3: histograms of the diversity measure in two (of 50) randomly selected data sets

Important Result

There is reasonable evidence that diversity at the district level as we have measured it seems to have a negative relationship with graduation rates.

Mathematical Section

We attempt to construct a general measure of diversity that will be high for racially diverse communities and low for racially homogeneous communities. By doing this, we hope to construct results that can be seen without reference to specific ethnic groups. We define the diversity statistic D_j for community j to be

$$D_j = \prod_{i \in I} (1 - x_{i,j}), \tag{1}$$

where I is an indexing set for ethnic groups and $x_{i,j}$ is the proportion of ethnic group i in community j. We want to determine whether D_j has a statistically

Prediction

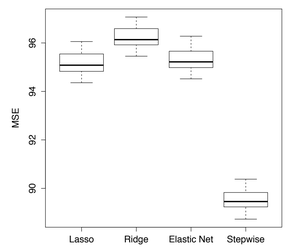


Figure 4: Comparison of Mean Squared Errors

Table 1: Penalization values

	λ_{min}	$\text{Var}(\lambda_{min})$	λ_{1se}	$\text{Var}(\lambda_{1se})$
Lasso	0.04595	0.00006	0.27143	0.00505
Ridge	1.46487	0.03459	6.77670	1.87964
Elastic Net	0.08589	0.00025	0.54794	0.01765

Inference

The below table details the significance of our Diversity Statistic estimated coefficient for all our different regression models.

Table 2: Regression summary

Model	Mean	Variance	Std. Error	t value	p value
Linear 1	-59.97	0.65	1.39	-43.28	0.00
Linear 2	-53.36	0.81	1.36	-39.29	0.00
Linear 3	-69.56	2.92	2.48	-28.04	0.00
Linear 4	-59.33	0.57	1.38	-43.09	0.00
Linear 5	-75.43	1.19	2.14	-35.21	0.00
Linear 6	-85.41	3.46	3.01	-28.34	0.00

Conclusion

We can see that our diversity statistic does indeed have a negative impact on graduation rates at the county level. Given that our data is public domain, we suspect that the bidirectional-elimination or lasso model would accurately predict graduation rates in future years. Furthermore, our model could be improved with a time series analysis of graduation rates along with our covariates over a number of years.

Limitations

There are various potential issues with our methods and resulting conclusion, including:

- Ecological Inference Problem
- Omitted Variable Bias
- Data Imputation Methods

Acknowledgements

I would first like to thank AT&T and the Data for Diplomas project for providing such a rich and interesting dataset to work on. I would like to acknowledge Professor Sayan Mukherjee for challenging me with my methods and results. I would also like to thank Sunith Suresh, Azeem Zaman, and Lin Xiao, with whom I worked on part of this project with in our Predictive Modeling course.

Contact Information

- Email: sanjay.hariharan@duke.edu
- Phone: +1 (414) 828 3096

Duke UNIVERSITY