The Effect of Racial Diversity on Graduation Rates in U.S. School Districts

Sanjay Hariharan
Department of Statistics
Duke University
sanjay.hariharan@duke.edu

Abstract

In this paper we assess whether a statistic designed to measure the holistic level of racial diversity in a school district has a statistically significant impact on the graduation rate in the school district. The analysis is performed using data from the Data for Diplomas project, which is an amalgamation of data from various sources, including the 2010 US census and the American Community Survey.

1 Introduction

In the education literature there are various theories about the effect of classroom composition [1] on education outcomes, such as graduation rate or test scores. One paradigm for measuring classroom composition is racial diversity. We are interested in exploring the relationship between racial diversity and education outcomes[2].

In this paper, we propose a measure of racial diversity and attempt to explore the relationship between this statistic and graduation rates across school districts in the United States. We expect a small but significant relationship between these two factors. In our analysis, we explore the relationship between these factors, devoting special attention to controlling for potentially confounding economic variables.

1.1 Data Cleaning

This data set was very challenging to work with and exhibited many problems common to large data sets. The data set was created by merging together information on high schools by school districts along with census information. Census information is not classified by school districts, so the two dataset were matched by degree of geographic overlap. The most challenging problem with the data set was missing data. We used the VIM packages in R to assess the missing data. We found that numerous feature columns had missing data in different proportions. We decided to drop observations with missing values for graduation rates (the response variable in our regression). Feature columns with small amount of missing data ($< 2\%$) were assumed to be 0.

Unfortunately, the columns with the largest quantity of missing data were the features that contained information on the racial composition of the students in the school district. This posed a serious problem, as we used this information to compute the diversity statistic. Since the diversity statistic is the central predictor in our model, omitting all observations with missing data is simply not possible; this would result in a loss of approximately a two-thirds of our data. Since we had information on the racial composition of the district on the population level from the census data, we decide to impute the missing data on racial composition of the students using multivariate imputation by chained equations (MICE)

implemented by the `mice` package in R [3]. We created 50 different datasets with imputed values. We were glad to find that the diversity statistics was fairly stable under all the datasets (see Figure 2). This approach relies on the assumption that the racial composition of the school district is similar to the racial composition of the community at large, which may not be entirely accurate, given the structure of the data set.

2 Understanding Diversity

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

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

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 significant impact on graduation rates once other factors are accounted for.

The use of such a statistic may be desirable due to the correlation between ethnicity and wealth. Since wealth is highly correlated with positive academic outcomes, there is a concern that the ethnic composition of a community maybe serving as a proxy for wealth in our analysis. We want to establish whether the level of diversity (measured using D_j) is significant after wealth and other demographic information has been taken into account. Median income and poverty level are among the first covariates we control for in our analysis.

3 Inference

In order to assess the relationship between racial diversity and graduation rates we ran regression models of varying complexity. Due to the high level of missing data, we created 50 different imputed data sets (see Section 1.1). The regressions are performed across all data sets to account for variability. Table 1 presents relevant summary statistics for the estimated coefficient for Diversity. These values are averaged over all 50 imputed data sets.

Table 1: 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
Lasso 1	-45.04	2.18	1.47	-30.57	0.00
Lasso 2	-62.23	499.20	20.37	-3.12	0.03
Lasso 3	62.38	530.68	22.11	2.87	0.06

In the linear models 1-6, we attempt to control for economic factors that correlated with diversity. We note however, that we only have information of economic factors at the district level. It maybe preferable to include information for racial groups within the district to entirely control for economic factors. However such information is not available in our data set.

In all models, diversity is found to be significant, as evidenced by the small p-values on the coefficient estimate. For the relatively simple models, Linear models 1-6 and Lasso 1, the estimate is consistently negative with a low variance, which suggests stability. In Lasso model 3, after adding in squared predictors, we find an interesting result of a positive coefficient estimate. However we also find large variances in our estimates for Lasso 2

Table 2: Description of Regression Models

Model	Covariates
Linear 1	Regression with diversity as the only covariate
Linear 2	Regression with diversity and median income in the district
Linear 3	Regression with diversity, median income in district, and their interaction
Linear 4	Regression with diversity and the number of people classified as being below the poverty level
Linear 5	Regression with diversity, poverty level, and their interaction
Linear 6	Regression with diversity, poverty, income, and their interactions with diversity
Lasso 1	Unpenalized regression on variables selected by lasso with the 1 squared error λ
Lasso 2	Unpenalized regression on variables selected by lasso (1se) with interaction terms
Lasso 3	Unpenalized regression on variables selected by lasso (1 se) with interaction between all variables and squared terms for each variable

and Lasso 3 models. These results suggest instability in the estimates from these models, possibly due to instability in lasso selection methods or increased collinearity from the added interaction and squared terms. We note that t values tend to decrease as model complexity increases.

Nevertheless, the t statistics and p-values remain highly significant even when other explanatory variables are included in the model. This is reasonable evidence that diversity at the district level as we have measured it seems to have a negative relationship with graduation rates. The quality of inference could perhaps be improved by using a more robust model, such a mixture model.

3.1 Predictions

We experimented with a variety of different models to assess the best predictive ability with our data. The total rate of graduation within each school district is our response, with all the other covariates in our dataset fixed. This section provides an overview of the different models we used and our rational behind them.

The first method of model selection we used was lasso. Given that we have over 350 covariates, the possibility of linear dependence and overfitting is quite high when performing full linear regression. The Lasso of Tibshirani (1996)[4] provides a method of variable selection, penalizing covariates and pushing their estimated coefficients to zero. We chose the penalization term λ through Cross-Validation, picking the value that minimizes out of sample deviance.

To provide an alternative to our Lasso model, we chose to model the response using Ridge Regression [5], which minimizes the covariate estimated coefficients, but does not zero any out. Like above, we chose the penalization term that minimizes the out of sample deviance.

Since the Lasso Model pushes many coefficient estimates to zero, and the Ridge model keeps all the estimates, we want to compare them against a model that is not so extreme, and rather weighs them equally. The Elastic Net Model does just this, weighing the Ridge and Lasso Regression Penalizations by a value α, and choosing the penalization term that minimizes the out of sample deviance [6]. For our purposes, we chose an equal weight, ie: $\alpha = 0.5$.

As an substitute for penalized regression, we chose a combination of Forward and Backward Stepwise Selection, called Bidirectional Elimination[7], and used AIC[8] as the selection criterion. This method iteratively adds in covariates that best improve the model fit. As covariates are added, the algorithm also removes previous variables that become insignificant compared to the new additions [9].

3.2 Result

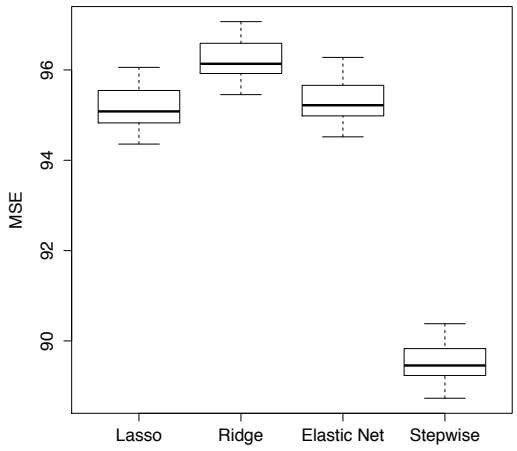

Figure 1: Comparison of Mean Squared Errors

Data Cleaning with the Mice Package in R resulted in 50 different imputed datasets. Through Cross-Validation, we implemented the four models above on each resulting dataset and computed the Mean Square Error of the model on the known response values. A box plot of the results is presented in Figure 1. The mean and variance of the shrinkage parameter across the 50 data sets are presented in Table 3

We can clearly see that the bidirectional elimination model fits our data set the best, followed by the lasso model. This result is due to the fact that both these models completely ignore covariates that do not significantly affect the model.

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.

Table 3: 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

4 Conclusion

Despite our best efforts to conduct a rigorous and thorough analysis, there are various potential problems. One area that may be especially problematic is our data cleaning methods. We began with a very messy data set and had to impute for large amounts of missing data.

The Ecological Inference Problem is relevant to our analysis. The aggregation of data across districts will tend to amplify the effects of the covariates on the response. That is to say, the effect of diversity is likely stronger on a district level than it is on a school level. Our data set does not include information on the number of high schools in each district, so we cannot restrict our attention to those districts with only one high school. An analysis on the school level might give drastically different results.

We may not have sufficiently controlled factors related to diversity. It is possible that important covariates were lacking and our estimate for the effect of diversity suffers from

omitted variable bias. Moreover, the effect of diversity may be regional in a manner that cannot be controlled by state. In particular, the attitudes towards racial diversity may differ between the Northern and Southern United States due to historical factors.

We cannot offer an explanation for the mechanism through which diversity affects graduation rates. We reiterate that our analysis was performed at the district level subject to various limitations. Nonetheless, we hope that the results in this paper provide an interesting and reasonable motivation for further study.

5 Appendix

6 Data

We have obtained data from the Data for Diplomas project. This project released a large data set containing Census, ACS Survey, and education specific data across over ten thousand school districts in the United States [10]. Information on Idaho, Oklahoma and Kentucky are absent from the data set. There are over 350 features in the dataset. We are primarily interested in data regarding overall graduation rate and racial composition in each school district.

As we only have data at the district level, it is difficult to make claims about the relationship between diversity and graduation rate in a school because school districts can have multiple high schools, which poses a problem for inference. It could be the case that a racially diverse district contains racially homogeneous schools. In our data set such an observation would appear as a single diverse district. This type of confounding will tend to strengthen the relationship between diversity and graduation rate, making inference difficult. This problem is exacerbated even more by the fact that students are not randomly assigned to schools. Instead, the majority are assigned to a high school based on residence. As ethnic groups commonly live in close-knit communities, there are likely to be high schools dominated by a distinct racial group [11].

The data set is rather messy and contains many missing values. A discussion of the procedures used to clean the data and impute missing values can be found in the Appendix in Section 1.1

6.1 Mathematical Properties of the Diversity Statistic

Let I be an indexing set containing various ethnic groups and let $x_{i,j}$ be the proportion of the population in community j that belongs to ethnic group i. We define the diversity statistic D_j for community j to be

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

This is subject to the constraints $x_{i,j} > 0$ for all i and j and $\sum_{i \in I} x_{i,j} = 1$. We will show that subject to these constraints the maximum value of D_j occurs when $x_{i,j} = 1/|I|$ for all i. We will begin with the more general constraint $\sum_{i \in I} x_{i,j} = c$ with $c \leq 1$. Fix some j. The Lagrangian is

$$L(\boldsymbol{x}, \lambda) = \prod_{i \in I} (1 - x_i) + \lambda \sum_{i \in I} x_i - \lambda c.$$

This results in first order conditions of the form

$$\frac{\partial L}{\partial x_i} = - \prod_{j \in I - \{i\}} (1 - x_j) + \lambda.$$

First suppose that $x_j = c$ for some $j \in I$. In this case all $x_i = 0$ for all $i \neq j$ and the value of the objective function is $1 - c$. Note that if $x_i = x_j = c/2$ the value of the objective function is $(1 - c/2)^2 > 1 - c$. Therefore the maximum does not occur when only one x_i is non-zero and we can restrict our attention to situations when $x_i < c$ for all all $i \in I$; in

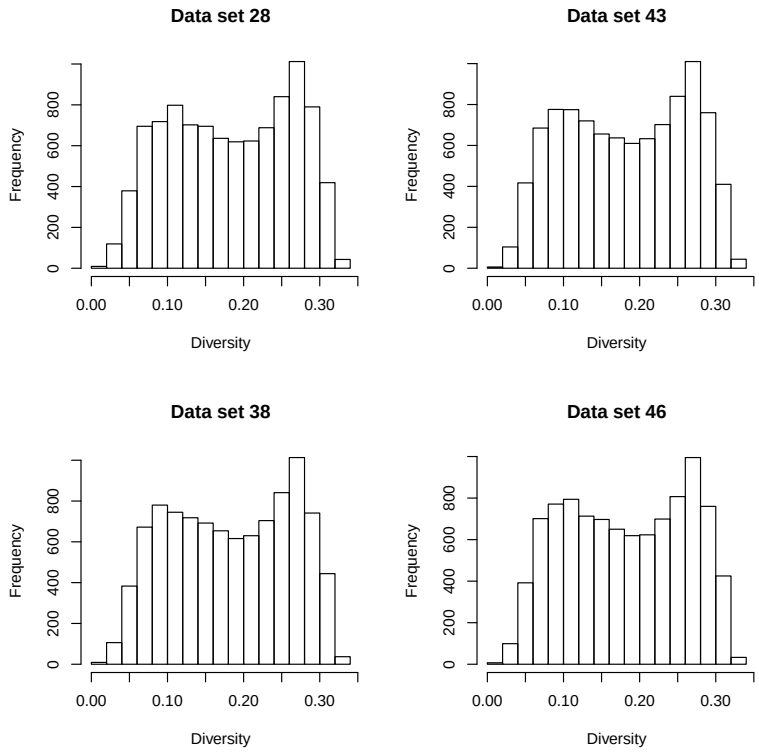

Figure 2: histograms of the diversity measure in four (of 50) randomly selected data sets

particular $x_i < 1$. This allows us use the equations derived from the Lagrangian to conclude that $x_i = x_j$ for all $i, j \in I$. Using the constraint on the sum we immediately find that $x_i = c/|I|$ for all $i \in I$.

In other words, the value D_j is maximized when all ethnic groups are equally represented in a community. Conversely, if there is a community j with only one ethnic group, then $D_j = 0$, which is the minimum possible value for D_j. Above we used the more general constraint with c to illuminate the following: if one racial group makes up some proportion of the community, the diversity statistic is maximized when the remaining ethnic groups are equally represented. These are appealing properties for a measure of diversity.

6.2 Empirical Stability of the Diversity Statistic

We have proposed a novel measure of diversity and are concerned that the measure may exhibit undesirable behavior. It would be desirable to have a measure that is stable under small deviations. We have generated data sets by imputing many values. The variability in the imputed values will create small variations in the diversity measure, which we can use to examine the stability of the measure. First we compare the histograms of the diversity measure in four (of 50) randomly selected data sets, which is presented in Figure 2. Note that the histograms of the observed measures of diversity are similar across data sets. This suggests that our measure is relatively stable. We can also verify that the diversity measure for the same observation in two randomly selected data sets is similar, which is presented in Figure 3. This is done for two pairs of randomly selected data sets. The histograms are highly concentrated around zero. This further suggests that the diversity statistic is robust to small changes in the data.

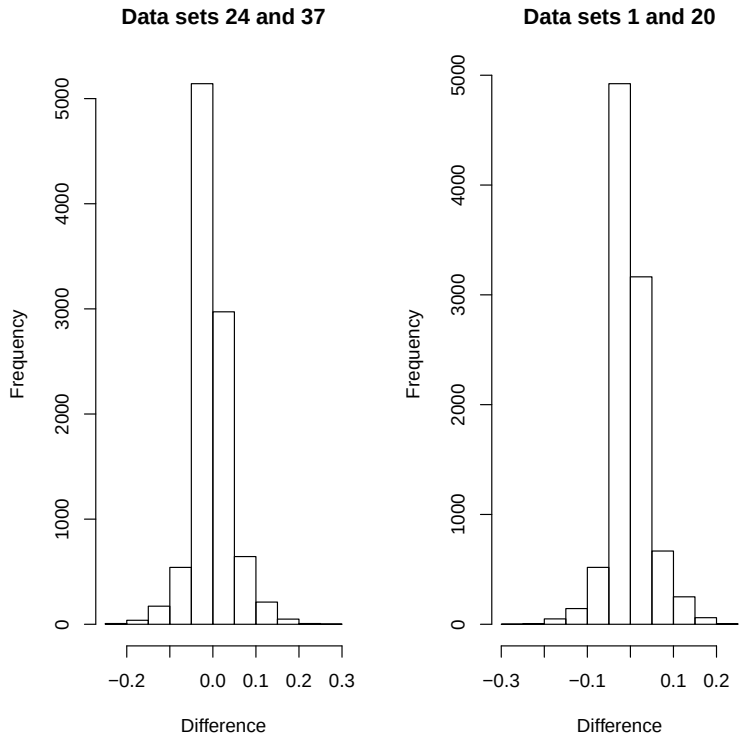

Figure 3: Diversity measure for the same observation in two randomly selected datasets

6.3 Visualing Diversity

Figures 4 and 5 show the value of the diversity statistic and graduation rate for school districts in the United States. The plot of the diversity measure has features that are intuitively appealing. For example, the Midwest has very low diversity as measured by our statistic. Many states in this region of the country are composed almost entirely of Caucasians, and the plot reflects this. Diversity is also higher in certain urban areas, such as Chicago. Furthermore, southern states, such as California or Arizona, where a large number of hispanics reside, are observed to have high diversity.

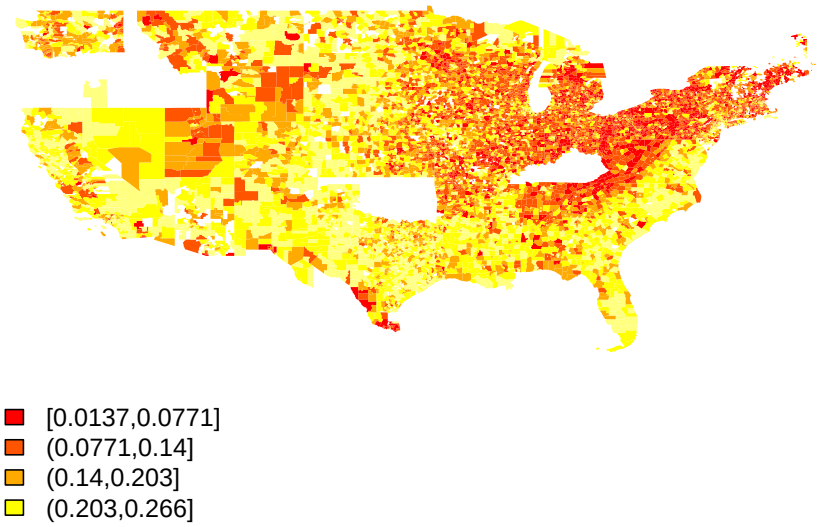

Figure 4: Plot of diversity in U.S by school district

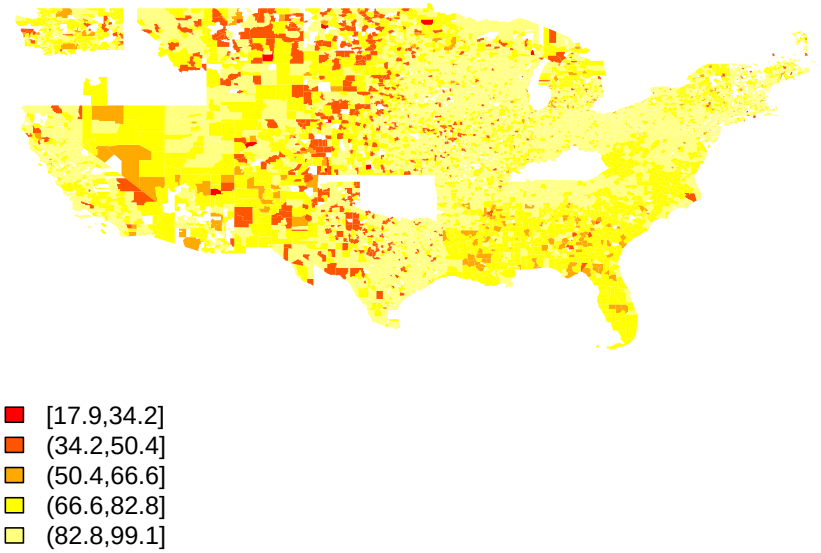

Figure 5: Plot of graduation rate in U.S by school district

References

[1] Amita Chudgar, Thomas F. Luschei, and Yisu Zhou. Science and Mathematics Achievement and the Importance of Classroom Composition: Multicountry Analysis Using TIMSS 2007. *American Journal of Education*, 119(2):295–316, Feb 2013.

[2] Yehezkel Dar and Nura Resh. Classroom Intellectual Composition and Academic Achievement. *American Educational Research Journal*, 23(3):357–374, 1986.

[3] Stef van Buuren, Karin Groothuis-Oudshoorn, Alexander Robitzsch, Gerko Vink, Lisa Doove, and Shahab Jolani. *mice: Multivariate Imputation by Chained Equations*, 2015. R package version 2.25, `http://CRAN.R-project.org/package=mice`.

[4] Robert Tibshirani. Regression Shrinkage and Selection via Lasso. *Journal of the Statistical Society. Series B(Methodological)*, 58(1):267–288, 1996.

[5] Donald W. Marquardt and Ronald D. Snee. Ridge Regression in Practice. *The American Statistician*, 29(1):3–20, 1975. `http://www.jstor.org/stable/2683673`.

[6] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005. `https://web.stanford.edu/~hastie/Papers/elasticnet.pdf`.

[7] Brian Ripley, Bill Venables, Douglas M. Bates, Kurt Hornik (partial port ca 1998), Albrecht Gebhardt (partial port ca 1998), and David Firth. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, 2015. R package version 7.3-45, `http://CRAN.R-project.org/package=MASS`.

[8] Hirotugu Akaike. A New Look at the Statistical Model Identification. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 19(6):716–723, Dec 1974.

[9] Wikipedia. Stepwise regression — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Stepwise_regression`, 2015. [Online; accessed 14-November-2015].

[10] Jennifer L. DePaoli, Joanna Hornig Fox, Erin S. Ingram, Mary Maushard, John M. Bridgeland, and Robert Balfanz. Building a Grad Nation(Progress and Challenge in Ending the High School Dropout Epidemic). May 2015. Instituitions: Civic Enterprisesa and Everyone Graduates Center at the School of Education at Johns Hopkins University.

[11] Robert Dreeben and Rebecca Barr. Classroom Composition and the Design of Instruction. *Sociology of Education*, 61(3):129–142, Jul 1988.