Who Can Revive American Dream: Predicting Intergenerational Mobility in the United States

Xiling (Celia) Zhu
xiling@uchicago.edu
The University of Chicago
Chicago, IL 60615

Abstract

This project surveys what neighborhood characteristics are important in predicting intergenerational income mobility in the United States. Intergenerational income mobility is measured by the income rank of American adults who were born to parents in the 25th percentile of national income distribution. I trained three models I learned from this course: Least squares, Truncated SVD, and Ridge regression, and used the cross validation method to assess out-of-sample accuracy. I found that employment rate, poverty rate, and elementary school quality of a neighborhood are most important neighborhood characteristics in predicting intergenerational mobility in the U.S.

1 Introduction

Intergenerational mobility, defined as the movement within or between social classes and occupations occurring from one generation to the next, has been one of the main concerns for many policymakers and individuals. In this project, I focus on predicting the chances of upward mobility for children from low-income households. I define the key measurement of intergenerational mobility as the mean percentile rank in the national distribution of household income in 2014-2015 for individuals with parents at the 25th percentile of the national income distribution in each county. The goal of this project is to find the best predictor of intergenerational mobility for children coming from low-income households.

2 Literature review

A canonical measure of intergenerational income mobility is the intergenerational elasticity of income (IGE), estimated by regressing the log income of an adult on log income of his or her parent(s) (Solon, 1992). However, this measure was found to yield very unstable estimates of intergenerational mobility, primarily because the relationship between log child income and log parent income is non-linear (Chetty, Hendren, Kline, and Saez, 2014). They proposed a measure based on the joint distribution of parent and child income,[1], which can be decomposed into two components: (i) the joint distribution of parent income and child income ranks, formally known as the copula of the distribution, and (ii) the marginal distributions of parent and child income. The marginal distributions determine the degree of inequality within each generation, typically measured by Gini coeffcients or top income shares (Chetty, Hendren, Kline, Saez, and Turner, 2014). The degree of inequality is

[1] In a two-generation setting, “child income” refers to the income of the second generation in their adulthood, and “parent income” refers to the income of the first generation.

relevant in measuring intergenerational mobility because income mobility and inequality are often negatively correlated (Krueger, 2012).

While the parents income rank is important in predicting child income rank, a wealth of literature explored how neighborhoods in which children grow up shaped their economic opportunities in adulthood and affected intergenerational mobility in the U.S (e.g., Durlauf, 1994; Katz, Kling, and Liebman, 2001). It was found that neighborhoods associated with high income upward mobility rate usually have less income inequality, less segregation based on race or poverty, better school quality, and stronger social network and community involvement[2] (Chetty, Hendren, Kline, and Saez, 2014).

3 Preliminaries

Data I used data from the *Opportunity Atlas* constructed by Chetty et al. in 2020. The data set records American adults' earnings, aggregated by the county in which they *grew up*, rather than the county they currently live in as adults. The data links with federal income tax returns data from 2005-2015 to obtain information on their parents' earnings. The data also links with the 2005-2015 American Community Surveys (ACS) to obtain neighborhood characteristics. The data set I used is aggregated on the county level. It primarily focused on individuals born in the US in 1978-1983, or authorized immigrants who came to the U.S. in childhood in 1978-1983.

Label The label I am interested in is the mean percentile rank in the national distribution of household income in 2014-2015 for children with parents at the 25th percentile of the national income distribution. I am interested in children whose parents make up the bottom of national income distribution. When these children grew up, what are their earning outcomes? Did they move upwards in the national income distribution?

Feature I selected 10 features for prediction: 1) bankruptcies per 1000 adults in 2014, 2) violent and property crime rate, 3) mean household income in 2000, 4) median household income in 1990, 5) fraction of residents with a college degree or more in 2006-2010, 6) percentage of population below poverty line in 2006-2010, 7) average school district level standardized test scores in 3rd grade in 2013, 8) employment rate in 2000, 9) job density (in square miles) in 2013, and 10) percentage of children eligible for free lunch.

On average, the individuals in our sample attained a higher percentile rank in national income distribution than their parents. In figure 1, I report the distribution of the label – mean percentile rank in the national distribution of household income in 2014-2015. In table 1, I report the descriptive statistics of the label.

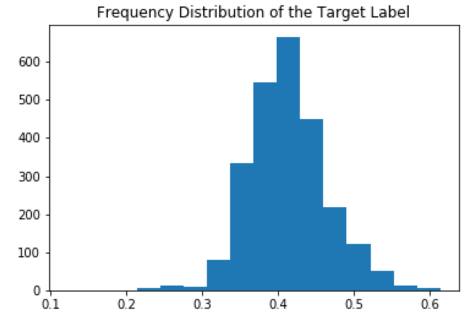

Figure 1: Distribution of mean percentile rank in the national distribution of household income in 2014-2015

[2] One proxy for social network and community involvement is the social capital index constructed by Rupasingha and Goetz (2008), which is comprised of voter turnout rates, the fraction of people who return their census forms, etc.

Table 1: Descriptive statistics of mean percentile rank in the national distribution of household income in 2014-2015

Label: mean percentile rank in the national distribution of household income in 2014-2015	
Statistics	Value
Sample size	2518
Mean	0.413
Standard Deviation	0.051
Min	0.122
25%	0.380
50%	0.411
75%	0.442
Max	0.614

Notations I use $\mathbf{X}$ to denote the training data, $\mathbf{y}$ label, $y_i \in [0, 1]$, $\hat{\mathbf{y}}$ predicted label, and $\hat{\mathbf{w}}$ predicted weight vector. In Truncated SVD, I use $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V}$ to denote the matrix whose columns are left singular vectors of $\mathbf{X}$, the matrix of singular values, and the matrix whose columns are right singular vectors of $\mathbf{X}$, respectively, and σ_i to denote the i-th singular value. I use Err to denote average training error rates, $CVErr$ average testing error rates using k-fold corss validation.

4 Models and Validation

4.1 Least squares regression

$\mathbf{X}$ is a full rank matrix. Therefore, I can use the simple equation to calculate $\hat{\mathbf{w}}_{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ I then calculated $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}$ and compared it with $\mathbf{y}$. The average training error rate $Err_{ls} = .237\%$.[3]

4.2 Truncated SVD

Next, I calculated the error rates by truncated SVD. First, $\mathbf{X}$ was decomposed into the left singular vectors ($\mathbf{U}$), the singular values matrix ($\mathbf{\Sigma}$), and the right singular vectors ($\mathbf{V}$). In figure 2, I inspected the spectrum of $\mathbf{X}$ over the singular values.

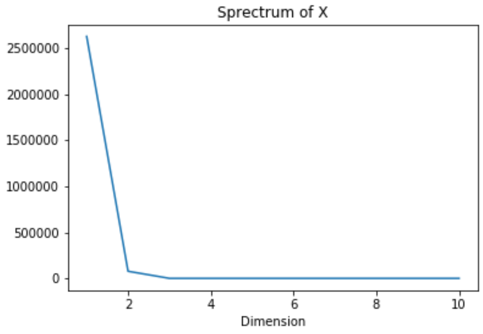

Figure 2: Spectrum of $\mathbf{X}$

The singular values in the 3 and higher dimensions are close to 0. This signals that the singular values important for Truncated SVD are only the first and second ones. Thus, I replaced the singular values other than σ_1 and σ_2 with 0, and calculated the pseudo-inverse matrix from the singular values. Then, I calculated $\hat{\mathbf{w}}_{svd} = \mathbf{V}\mathbf{\Sigma}^{+}\mathbf{U}^T\mathbf{y}$, and the predicted labels are $\hat{\mathbf{w}}_{svd} = \mathbf{X}\hat{\mathbf{w}}_{svd}$, the average training error rate of truncated SVD is $Err_{svd} = 0.790\%$.

[3] The formula I use to calculate average training error rates in this project is $\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2/N$.

4.3 Ridge regression

I calculated $\hat{\mathbf{w}}_r = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. The list of regularization parameter λ I used is $\{0.2, 2, 4, 8, 16, 32, 64\}$. The average training error rates, corresponding to each λ, are around 0.00789. Figure 3 reports the training error rates as I choose different regularization parameter. On average, $Err_r = 0.790\%$

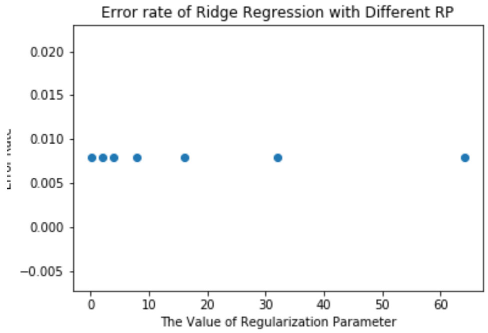

Figure 3: Error rate of Ridge Regression with different regularization parameter

4.4 Validation

Based on the average training error rate, least squares is the most accurate model. In this section of model validation, I focused on least squares model and used cross-validation to calculate the testing error rate of the least squares model.

I divided the data set into 26 groups. Each group has 100 observations. Each time, one group will be treated as a testing group, and the rest of groups will be training groups. I calculated the weight from training groups and applied that weight to the test group and got. Finally, I took the average of the error rate of these 26 loops I did. The average testing error rate is $CVErr_{ls} = 0.276\%$, which is still smaller than the training error rate of the SVD and the Ridge regression. Therefore, among the three models I trained, the least squares model is the most accurate one.

Moreover, I compared the absolute values of the weights to identify the significant predictors in the least squares model. The top 3 significant predictors are the employment rate in 2000, the percentage of residents below poverty line in 2006-2010, and the average school district level standardized test scores in 3rd grade in 2013.

5 Conclusion

The results from the least squares model indicate that the most important predictors for intergenerational mobility is not household income. Rather, the employment rate has the highest weight in predicting intergenerational mobility. Also, quality of elementary school education and poverty rate are significant predictors.

Standardized test scores of elementary schools and poverty rate are traditional proxies for neighborhood disadvantage. Our results showed that neighborhood disadvantage does predict intergenerational mobility. But that does not capture all variance.

Employment rate is the most important predictor in the model, while job density is less important. Evidently, what better predicts intergenerational mobility is not the job availability in the neighborhood, but “the company you keep” – whether children grow up with people who have jobs.

References

[1] Krueger, A. B. (2012) The Rise and Consequences of Inequality in the United States.

[2] Solon, G. (1992). Intergenerational income mobility in the United States. *American Economic Review*, 82(3), 393–408.

[3] Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States. *Quarterly Journal of Economics*, 129(4), 1553–1623.

[4] Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States still a land of opportunity? Recent trends in intergenerational mobility. *American Economic Review*, 104(5), 141–147.

[5] Durlauf, S. N. (1994). Spillovers, stratification, and inequality. *European Economic Review*, 38(3–4), 836–845.

[6] Katz, L. F., Kling, J. R., & Liebman, J. B. (2001). Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment. *The Quarterly Journal of Economics*, 116(2), 607–654.