The Factors Impacting the American Intergenerational Stratum Mobility

Yao Weitong

The idea of this topic comes from *Race and Economic Opportunity in the United States: An Intergenerational Perspective*[1], written by Raj Chetty, Nathaniel Hendren, Maggie R. Jones, Sonya R. PorterIs, Harvard University. The research has been published on New York Times and it ephasized how difference of the stratum mobility between black males and white males. They found the intergenerational persistence of disparities varies substantially across racial groups and black Americans have substantially lower rates of upward mobility and higher rates of downward mobility than whites. In this assignment, I use GSS data to study the factors impacting the American intergenerational stratum mobility, considering all responders of the interview as a whole sample instead of focusing their sexes and colors. ***Github Page***

1 Review of Ridge, Lasso and Elastic Net

Ridge Regression

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - BX_i)^2 + \lambda||B||^2 = MSE(\theta) + \lambda\sum_{i=1}^{m}\theta_i^2$$

$\sum\limits_{i=1}\theta_i^2$ is called L2 regularizer. In R, we can use the package *ridge* and method *linearRidge(.)*.

Lasso Regression

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - BX_i)^2 + \lambda||B||^2 = MSE(\theta) + \lambda\sum_{i=1}^{m}|\theta_i|$$

$\sum\limits_{i=1}|\theta_i|$ is called L1 regularizer. In R, we can use the package *glmnet* and method *glmnet(.)* for Lasso and Elastic Net.

Elastic Net

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - BX_i)^2 + \lambda||B||^2 = MSE(\theta) + \lambda\sum_{i=1}^{m}\Omega(\theta)$$

where $\Omega(\theta) = \alpha L_1 + (1-\alpha)L_2$ is the combination of L1 and L2 regularizer.
These three regression estimates are not scale equivariant so X needs to be standardized first. The processs chooses the optimal λ is the process of model selection. Differences are intuitionistic, different regularizers, for Ridge, the L2 regularizer allows the shrinkage of all coefficients but not elimination; for Lasso, it allows some of coefficients towards zero but sometimes it will drop the coefficients randomly; for Elastic Net, it has combined the advantages of first two.

2 Data Preprocessing and Variable Selections

SEI is the abbreviation for Socioeconomic Index, reflecting the education, income, and prestige associated with different occupations[4]. This paper has defined the intergenerational stratum mobility as the gap between the responder's SEI and weighted average of parent's SEIs. If the gap is greater than 30, it indicates an upward mobility (defined as Y = 1); the downward stratum mobility occurs if the gap is smaller than -20 (defined as Y = -1), and if the stratum doesn't change, it's defined as Y = 0.
According to the previous study, Rothstein and Wozny (2012) found parental income mattered in stratum mobility when it came to racial disparity. Also, the environmental factors are important, for example, education (Dobbie & Fryer, 2011), sextual and racial discriminations (Bertrand & Mullainathan, 2004), the crime rate of the born place and the neighborhoods (Smith,2005).

personal information	Sex	degree	Degree
	race(race of respondent)		Spdeg (spouse's highest degree)
	hhrace (race of household)		Colsci(r has taken any college-level sci course)
	wkracism (r feels discriminated because of race)	parental edu	padeg (father's highest degree)
	sei (respondent socioeconomic index (1980))		Madeg (mothers highest degree)
	Rincome (respondents income)		Paeduc (highest year school completed, father)
	sei10(r's socioeconomic index (2010))		Maeduc (highest year school completed, mothe)
	parsol (r's living standard compared to parents)	dependent variable	Finrela (opinion of family income)
	rank(r's self ranking of social position)		incom16 (r's family income when 16 yrs old)
	incdef (distance below poverty line)		income06(total family income)
	Realrinc (r's income in constant $)		pasei10(r's father's socioeconomic index (2010))
	Realrinc (r's income in constant $)		masei10(r's mother's socioeconomic index (2010))
	abpoor (low income--cant afford more children)	environment	Affrmact (avor preference in hiring blacks)
	income(total family income)		Wrkwayup
	Kidssol (r's kids living standard compared to r)		racdif1 (differences due to discrimination)

family	granborn(how many grandparents born outside u.s.) Parborn (were r's parents born in this country) Granborn (how many grandparents born outside u.s.) Ethnic (country of family origin) family16 (living with parents when 16 yrs old) coninc (family income in constant dollars) wayraise（how likely situation caused by the way raised） dwelown16（did rs family own or rent home at age 16）	job	wrkgovt (govt or private employee) partfull(was r's work part-time or full-time?) satjob1 (job satisfaction in general) joblose(is r likely to lose job) Mustwork (mandatory to work extra hours) Jobinc (high income) Jobsec (no danger of being fired) Jobmeans (work important and feel accomplishment)
social relations	raclive (any opp. race in neighborhood) socrel (spend evening with relatives) socommun (spend evening with neighbor) socfrend (spend evening with friends)	personal quality	Life (is life exciting or dull) sprtprsn(r consider self a spiritual person) getahead(opinion of how people get ahead) Learnnew (job requires r to learn new things)
religion	denom16 (denomination in which r was raised) relig16 (religion in which raised)	born place	res16 (type of place lived in when 16 yrs old) reg16 (region of residence, age 16)

Based on the available GSS data, this paper chooses 60 related variables collected from 2000 to 2018, including four variables for dependent variables calculation and adding some variables reflecting personal quality. These variables can be divided into 11 categories: personal information, education[2], parents' educations, parental income, family, environments, born place[3], job, social relations, individual quality and religion. Most of these variables are dummy except for some income variables. Here are some statistical outcomes:

race = white,	sex = male				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	8,966	45.4	21.94473	9	92.8
race = white,	sex = female				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	10,439	39.7	22.59543	9	93.7
race = black,	sex = male				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	1,458	32	19.45902	10.6	92.8
race = black,	sex = female				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	2,291	33.2	20.93098	10.6	92.8
race = other,	sex = male				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	1,133	36.4	22.94537	12.6	92.8
race = other,	sex = female				
Variable	Obs	Median	Std. Dev.	Min	Max
SEI	1,157	37.6	1.270243	10.6	92.8

Variable	Obs	Mean	Std. Dev.	Min	Max
year	26,698	2008.533	5.756568	2000	2018
id	26,698	1427.166	947.2628	1	4510
wrkgovt	25,257	1.811221	.391341	1	2
paeduc	18,436	11.56558	4.196875	0	20
maeduc	22,087	11.58129	3.736304	0	20
degree	26,662	1.583865	1.207387	0	4
padeg	19,219	1.142567	1.233938	0	4
madeg	22,942	1.078241	1.080649	0	4
spdeg	11,499	1.707453	1.246316	0	4
sex	26,698	1.553712	.497116	1	2
race	26,698	1.337703	.6414083	1	3
res16	25,151	3.593456	1.526077	1	6
reg16	26,698	4.370964	2.709133	0	9
family16	25,172	2.057882	1.888563	0	8
incom16	20,790	2.741607	.9261517	1	5
parborn	25,099	1.266903	2.824431	0	8
granborn	23,737	1.175844	1.625183	0	4
income	23,341	10.87374	2.362073	1	12
rincome	15,911	10.22997	2.952564	1	12
income06	11,524	16.71598	5.753774	1	25
relig16	25,040	1.941174	1.797247	1	13
denom16	13,794	35.29948	20.57787	10	70
raclive	21,274	1.280107	.4490619	1	2
affrmact	14,037	3.200613	1.013484	1	4
wrkwayup	14,642	2.153531	1.259154	1	5
life	14,629	1.550277	.585946	1	3
socrel	14,827	3.373103	1.650127	1	7
socommun	14,829	4.646841	2.02772	1	7
socfrend	14,828	3.951173	1.610368	1	7
partfull	17,723	1.216837	.4121022	1	2
joblose	9,405	3.488038	.7970266	1	4
jobinc	3,679	2.495787	1.205805	1	5
jobsec	3,678	3.444807	1.289093	1	5
jobmeans	3,687	2.214809	1.317056	1	5
rank	15,084	4.739592	1.824442	1	10
finrela	22,072	2.850761	.8938598	1	5
getahead	15,423	1.425274	.6802143	1	3
parsol	14,724	2.255569	1.144854	1	5
kidssol	14,600	2.71863	1.606296	1	6
abpoor	14,214	1.556564	.4968077	1	2
racdif1	14,208	1.613739	.4869088	1	2
wayraise	2,492	2.599518	.9116577	1	4
sprtprsn	16,560	2.130193	.9516769	1	4
mustwork	7,232	1.726493	.4457895	1	2
learnnew	7,313	1.736633	.7672841	1	4
wkracism	8,449	1.946384	.2252713	1	2
satjob1	7,312	1.655088	.7413506	1	4
hlthmntl	2,329	2.337484	.971841	1	5
colsci	8,845	1.578858	.4937702	1	2
realrinc	16,010	24911.28	37419.6	227	480144.5
coninc	23,673	49340.68	42789.05	350.5	178712.5
ethnic	21,232	17.91301	17.04089	1	97
dwelown16	1,550	1.287097	.4908734	1	3
hhrace	26,447	1.499338	1.111117	1	5
sei	15,967	49.43349	19.4774	17.1	97.2
sei10	25,444	45.78475	22.39835	9	93.7
pasei10	19,991	45.60974	20.42658	9	92.8
masei10	16,075	39.26247	22.54169	9	92.8

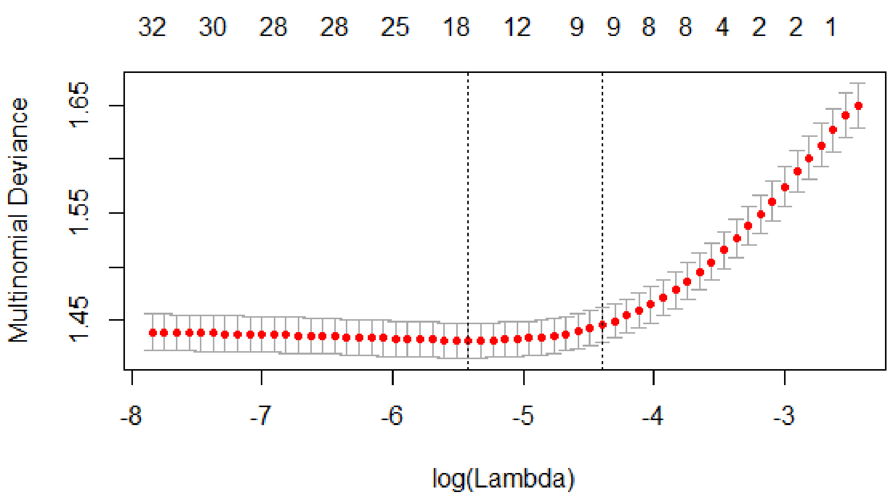
3 Model Selection and Elastic Net

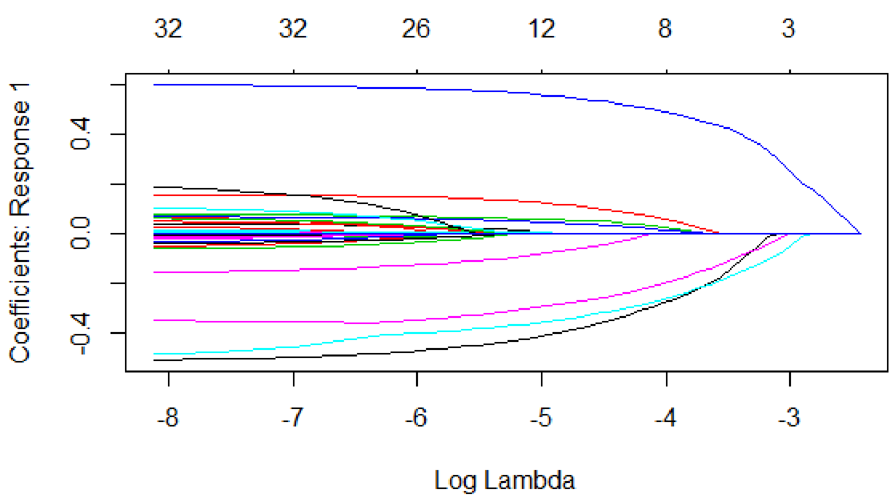
3.1 Elastic Net Regression using Cross Validation

One problem of using glmnet(.) for Lasso regression and Elastic Net regression is that it can't handle data with missing values. The first step is to eliminate the variables which have most missing values and least explanations to the dependent variables gap. Then using mean or mode(it depends on the nature of variables) of each id for each variable to fill the missing values, the data set with more than 30,000 records shrinks to 9591 records because of the serious missing values of SEI (dependent variables). All data preprocessing is done by Stata and regression is done by R.
Here are the results using Lasso regression:

```
Call:  glmnet(x = x1, y = y, family = "multinomial", alpha = 1, lambda = lambda.star)

      Df    %Dev   Lambda
[1,] 30 0.1448 0.004433
```





```
> coef(lasso_1.star)
$`-1`
41 x 1 sparse Matrix of
                     s0
           -1.288130e+00
wrkgovt     4.239361e-01
paeduc      .
maeduc      5.596925e-02
degree     -3.338465e-01
padeg       2.978385e-01
madeg       3.007739e-01
spdeg      -8.103851e-02
sex         1.014837e-01
race        .
res16       .
reg16      -3.310924e-03
family16    .
incom16     5.244417e-02
parborn     1.121452e-02
granborn    .
income      .
rincome    -1.561280e-02
income06    .
relig16     .
denom16     .
raclive     6.120364e-02
affrmact   -2.044775e-02
wrkwayup   -1.199036e-02
life        7.147179e-02
socrel      .
socommun    .
socfrend   -6.018536e-03
rank        3.694033e-02
finrela    -7.354992e-02
getahead    4.259397e-02
parsol      .
racdif1    -3.628143e-02
wayraise    .
sprtprsn    .
wkracism    .
colsci      1.087673e-02
realrinc   -5.066323e-06
coninc     -2.685027e-06
ethnic      .
hhrace      .
```

```
$`0`
41 x 1 sparse Matrix of
                     s0
             1.967126012
wrkgovt     .
paeduc      .
maeduc      .
degree      .
padeg       .
madeg       .
spdeg       .
sex         .
race        0.154232955
res16       .
reg16       .
family16    .
incom16     .
parborn     .
granborn   -0.044099281
income     -0.026507876
rincome     .
income06    .
relig16     .
denom16     .
raclive    -0.017667270
affrmact    .
wrkwayup    .
life        .
socrel     -0.018013725
socommun    .
socfrend    0.048680177
rank        .
finrela     .
getahead    .
parsol      .
racdif1     .
wayraise    0.016740921
sprtprsn    .
wkracism    .
colsci      .
realrinc    .
coninc      .
ethnic      0.001108856
hhrace      .
```

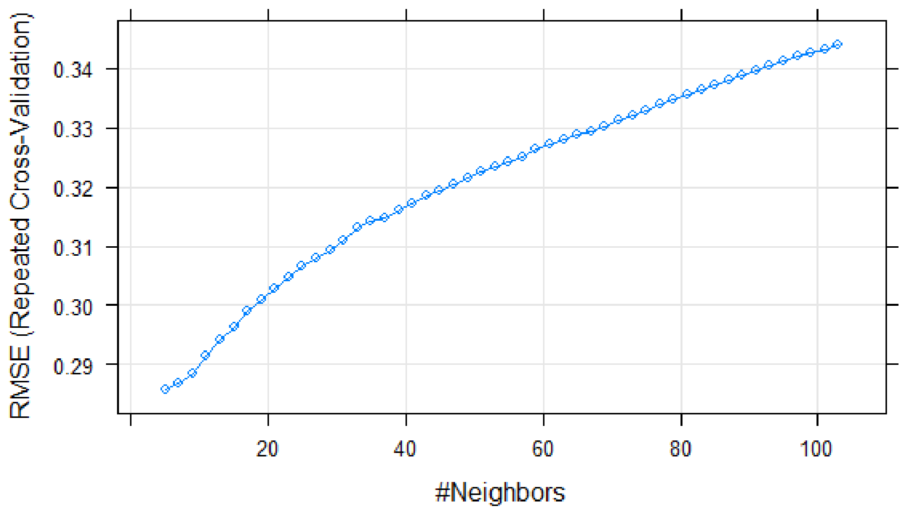
```
$`1`
41 x 1 sparse Matrix of
                     s0
           -6.789962e-01
wrkgovt    -4.453717e-01
paeduc      .
maeduc      .
degree      5.730284e-01
padeg      -3.814368e-01
madeg      -3.245803e-01
spdeg       1.738848e-02
sex         .
race       -8.927154e-03
res16       .
reg16       .
family16    .
incom16    -1.061561e-01
parborn    -8.553992e-03
granborn    .
income      6.236501e-02
rincome     5.306801e-02
income06    4.964345e-03
relig16     .
denom16     .
raclive     .
affrmact    1.265829e-03
wrkwayup    5.154645e-03
life        .
socrel      .
socommun    .
socfrend    .
rank        .
finrela     1.369034e-01
getahead   -1.412917e-02
parsol      .
racdif1     1.250114e-02
wayraise    .
sprtprsn    .
wkracism    .
colsci      .
realrinc    6.979121e-07
coninc      5.887341e-07
ethnic      .
hhrace      .
```

In Lasso regression, the data sets are seperated into training data(4795 obs.) and test data(4796 obs.). Using training data and cross-validation to minimize L1 error, glmnet can produce the optimal $\lambda = 0.004433$ and shrink some variables towards zero. The optimal λ should be calculated by cross-validation because Lasso needs a stable hyperparameter (the X used to produce optimal λ should not be highly correlated!).Then, use the testing data set to predict and evaluate this model:

```
> newx <-model.matrix(gap_up~.,GSS.te)[,-1]
> newx_1<-newx[,1:40]
> lassopred<-predict(lasso_1.star,newx_1,type="class")
> table(lassopred,GSS.te$gap_up)

lassopred   -1    0    1
       -1   55   20    0
       0   644 3244  633
       1     2   72  126
> lassoerr<- 1- mean(lassopred ==GSS.te$gap_up)
> lassoerr
[1] 0.2858632
```

If using KNN (K-Nearest Neighbour) to test the model on test data set, the result is as follows. The optimal λ does lead to a minimum error = 0.286.



3.2 Explanation of Results

Although the parameters predict Y=1 and Y=-1 are not all the same (treat Y = 0 as the reference level), the opposite sign of same coefficients do make sense. Take the probability of upward stratum mobility as an example, the economically significant parameters(e-02) are wrkgovt (=1:work for govt; =2: private), degree(0:lt high school - 4:graduate), padeg(dad's degree, same as label degree), madeg(mom's degree, same as label degree), spdeg(spouse's degree, same as label degree), income(total family income,;1-12, the larger the label is, the higher the income is), rincome, finrela, getahead and racdif1(difference due to discrimination; =1: yes, =2: no). All the coefficients appear to common sense but be slightly different from the results derived by Raj Chetty, Nathaniel Hendren, Maggie R. Jones, Sonya R. PorterIs, who foucus on racial perspectives and keep others constant. Instead of focusing on races, this paper takes all responders as entire sample and concludes that the generation stratum mobility depends on the degrees of two generations, and the parental income is quite significant. These two factors are also applied to compute SEI. The frictions derived from racial and sexual discrimination also impact the mobility at a statistically significant level.

References

[1] *Race and Economic Opportunity in the United States: An Intergenerational Perspective.* Raj Chetty, Nathaniel Hendren, Maggie R. Jones, Sonya R. PorterIs. Harvard University. March 2018.

[2] *Educational Progress For African Americans And Latinos In The United States From The 1950s To The 1990s: The Interaction Of Ancestry And Class.* Michael Hout. University of California, Berkeley. June 1999.

[3] *The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.* Raj Chetty, John Friedman, Nathaniel Hendren, Maggie R. Jones, Sonya R. Porter. Working Paper. October 2018.

[4] Occupational prestige, WIKIPEDIA.