Empirical Research by Lasso — using GSS data

Jiaqi Zeng 15220152202409

April 11, 2019

1 Review

Lasso means least absolute shrinkage and selection operator, wanting to find a *beta* to solve:

$$\hat{\beta}_{lasso} = \beta argmin \underbrace{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}_{SSR} + \underbrace{\lambda \|\beta\|_1}_{penalty}$$

As with ridge regression, the lasso shrinks the coefficient estimates towards zero. Lasso yields sparse models that is, models that involve only a subset of the variables.

2 Overview on Data and Variables

First of all, we choose education(highest year of school completed) as our target variable and 50 feature variables.

According to GSS, we choose feature variables by three ways:

1. The features of respondants.
2. The features of respondants' parents.
3. The features of environment respondants grow up.

So we choose 50 features to explain what affect respondants' education. Here are feature variables with descriptive summary.

Variable Name	Explaination	Variable Name	Explaination
educ	target variable	spaneng	spanish or english
born	was r born in this country	sex	respondents sex
denom	specific denomination	sexornt	sexual orientation
partyid	political party affiliation	fund16	how fundamentalist 16
coneduc	confidence in education	conschls	confidence in schools
inteduc	interested in local school issues	family16	how fundamentalist 16
teensex	sex in teens	jew16	jew 16
age	age of respondant	age2	age*age
sibs	number of brothers and sisters	evpaidsx	sex paid since 18
fatalism	can't change their lives	godchnge	beliefs about god
gunlaw	favor or oppose gun permits	health	condition of health
hell	belief in hell	hsbio	r took biology course
hschem	took chemistry course	hsmath	highest math in high school
evcrack	r ever use crack cocaine	evidu	r ever use crack inject drugs
unemp	ever unemployed in last ten yrs	maeduc	mother education
marelkid	mothers religion	masei10	mother's socioeconomic index
mawrkgrw	mothers employment	mawrkslf	mother work
paeduc	father education	parelkid	fathers religion
pasei10	father's socioeconomic	income16	family income 16
pawrkslf	father work	parborn	parents born
fampress	family pressure		
ethnic	country of family origin	res16	lived 16 yrs old
reg16	region of residence, age 16	relig16	religion in which raised
denom16	denominationd	oth16	other denominations
dwelown16	owned or rented home when 16	mobile16	geographic mobility

Then we see the summary of variables.

Variable	Obs	Mean	Variable	Obs	Mean
born	2347	1.128675	granborn	2214	1.183379
sex	2348	1.551959	denom	1152	42.12153
sexornt	1373	2.917698	partyid	2315	2.882505
fund16	2251	1.937805	coneduc	1545	1.937217
conschls	1163	2.846948	inteduc	1175	1.64766
family16	2347	2.214742	teensex	1534	1.752282
jew16	41	2.756098	age	2341	48.97138
age2	2341	2724.252	sibs	2343	3.579599
evpaidsx	1388	1.93804	fatalism	1165	4.027468
godchnge	1124	3.511566	gunlaw	1541	1.28488
health	1569	2.104525	hell	1142	1.916813
hsbio	1107	1.200542	hschem	1112	1.429856
hsmath	1058	4.608696	evcrack	1393	1.940416
evidu	1391	1.970525	unemp	1560	1.659615
maeduc	2089	11.87841	marelkid	1137	2.057168
masei10	1657	40.94315	mawrkgrw	2246	1.25423
mawrkslf	1671	1.897666	paeduc	1687	11.88322
parelkid	1068	2.530899	pasei10	1850	45.95643
respnum	2348	1.431857	mobile16	2348	1.987223
income16	2152	17.72862	parelkid	1068	2.530899
pawrkslf	1861	1.761956	parborn	2342	1.377028
fampress	1168	1.83476	ethnic	1833	18.86416
res16	2348	3.684412	reg16	2348	4.513629
relig16	2332	1.913379	denom16	1226	36.07341
oth16	237	56.05063	dwelown16	1550	1.287097

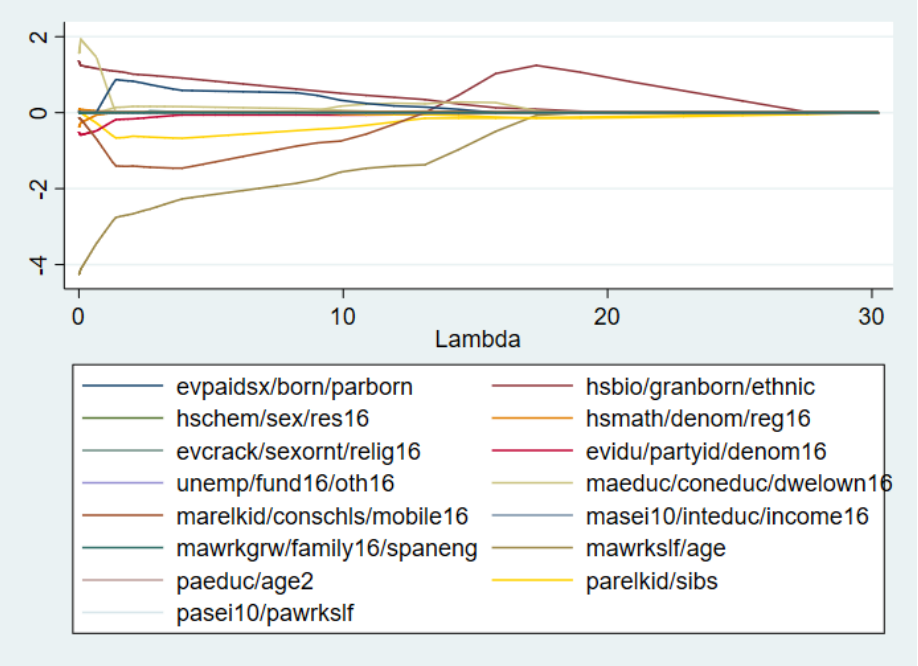
3 Select variables

Next, we select suitable variables using k-fold cross validation.

```
lasso2 educ evpaidsx hsbio hschem hsmath evcrack ///
>     evidu unemp maeduc marelkid ///
>     masei10 mawrkgrw mawrkslf paeduc parelkid pasei10 ///
>     born granborn sex denom ///
>     sexornt partyid fund16 coneduc conschls inteduc ///
>     family16 age age2 sibs pawrkslf ///
>     parborn ethnic res16 reg16 relig16 denom16 ///
>     oth16 dwelown16 mobile16 income16 ///
>     parelkid spaneng pawrkslf parborn ethnic ///
>     res16 reg16 relig16 denom16 oth16 ///
>     dwelown16 mobile16 income16 parelkid spaneng, plotpath(lambda)
```

Knot	ID	Lambda	s	L1-Norm	EBIC	R-sq	Entered/removed
1	1	30.24014	1	0.00000	28.81916	0.0000	Added _cons.
2	2	27.55369	5	0.08760	57.80051	0.0640	Added hsbio parelkid denom sibs.
3	6	18.99169	7	1.36488	69.45593	0.3002	Added granborn parborn.
4	7	17.30452	8	1.74823	75.93869	0.3577	Added mawrkslf coneduc. Removed parborn.
5	8	15.76724	10	2.20676	89.24534	0.4420	Added partyid parborn.
6	9	14.36652	11	2.32247	94.80645	0.5291	Added oth16.
7	10	13.09024	10	2.39052	85.54160	0.6015	Removed hsbio.
8	11	11.92734	12	2.94070	98.41521	0.6672	Added maeduc mobile16.
9	13	9.90228	11	3.89898	86.91589	0.7702	Removed oth16.
10	15	8.22105	9	4.54932	68.22053	0.8379	Removed coneduc sibs.
11	23	3.90566	10	6.13981	63.07552	0.9483	Added family16.
12	27	2.69202	11	6.68173	64.54028	0.9699	Added age2.
13	28	2.45287	11	6.72644	63.03112	0.9738	Added income16. Removed family16.
14	31	1.85551	12	6.96381	66.12407	0.9823	Added pasei10.
15	32	1.69067	11	7.02773	57.07235	0.9848	Removed income16.
16	35	1.27893	13	7.21113	67.30986	0.9900	Added coneduc reg16.
17	43	0.60759	11	7.70348	36.08271	0.9977	Removed maeduc parborn.
18	66	0.07150	12	8.39820	-3.54279	1.0000	Added maeduc parborn. Removed parelkid.

And we get a picture about solution path:



We can see the estimation of Lasso as follow:

Selected	Lasso	Post-est OLS
. hsmath	-0.0635505	
. maeduc	0.0753577	
. mawrkslf	-3.4403012	
. parelkid	-0.3176810	
. pasei10	0.0168219	
. granborn	0.6589222	
. denom	0.0519894	
. partyid	-0.4219740	
. coneduc	0.8784945	
. age2	-0.0002468	
. parborn	3.3832129	
. res16	0.2679737	
. reg16	-0.1697767	
. mobile16	-0.5710882	
Partialled-out*		
cons	18.3817811	

4 Linear Regression

We use linear model to estimate relationship between selected variables and education. Table1 shows the results of OLS (Turn to the last).

5 Conclusion

Lasso is a great method which helps us select variables and shrink some variables coef to zero in order to get a more suitable model.

References

[1] Tibshirani, The Lasso Page

[2] Chen,Econometrics and Application in Stata

Table 1: **Education**

	Group A
	(1)
	educ
hsmath	0.541***
	(4.64)
maeduc	0.158
	(1.33)
mawrkslf	−1.945
	(−1.65)
parelkid	−0.152
	(−1.46)
pasei10	0.0122
	(1.12)
granborn	0.504*
	(1.78)
denom	−0.00236
	(−0.19)
coneduc	−0.0316
	(−0.10)
age2	0.000295**
	(2.15)
parborn	−0.398**
	(−2.39)
res16	0.0550
	(0.30)
reg16	−0.166*
	(−1.89)
mobile16	0.430
	(1.43)
_cons	12.22***
	(4.45)
N	89

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$