hw07: PartII

Bohan Yin

11/17/2019

Contents

Predicting attitudes towards racist college professors

The General Social Survey (GSS) conducted the national survey with the question of "Should a person who believes that Blacks are genetically inferior be allowed to teach in a college or university?" Regarding this question, many explanations focus on a resource model of personal features – such as age, race, received education, income and etc. In its questionnaire, the respondents' were asked to respond to this question by filling out several columns of their personal information and thoughts, and according to the survey, 49% of Americans answered "Yes" to this question. How do those features affect the attitude towards racist college professors? How are they differ in terms of their influence?

The following analysis tries to investigate in attitudes towards racism, using the GSS and this specific question. It aims to find some of the underlying factors that affect the attitude toward racist college professors. To perform this analysis, I used two models, one logistic regression model and one model from lasso penalized regression, and compared their results trying to see if there is a better model to fit.

Model 1

Since it is not clear to decide what variables should select for this regression, the following analysis will compare two different models and check to see which one performs a better result. The first model is just a plain model, which includes all of the variables from the survey. The aic value reflects the predictive power of the model, which means that the lower aic value, the better the model is.

Here is the glimpse of model 1

Table 1: Logistic Regression Results

Variable	Coefficient	Std. Error	T-Statistic	P Value
(Intercept)	-6.9217	0.8302	-8.3377	0.0000
age	0.0085	0.0036	2.3751	0.0175
authoritarianism	0.0961	0.0412	2.3353	0.0195
blackYes	-0.0516	0.1852	-0.2785	0.7806
degreeHS	-0.7832	0.1926	-4.0671	0.0000
degreeCollege	-0.7994	0.2314	-3.4548	0.0006
degreeGraduate deg	-0.8943	0.2784	-3.2124	0.0013

Variable	Coefficient	Std. Error	T-Statistic	P Value
egalit_scale	-0.0022	0.0074	-0.3018	0.7628
grassNOT LEGAL	-0.0564	0.1505	-0.3747	0.7079
hispanic_2Yes	0.1625	0.1800	0.9026	0.3668
income06$1 000 TO 2 999	-0.2082	0.6815	-0.3055	0.7600
income06$3 000 TO 3 999	0.4576	0.7795	0.5870	0.5572
income06$4 000 TO 4 999	0.5344	0.8754	0.6105	0.5415
income06$5 000 TO 5 999	-1.3083	1.1731	-1.1152	0.2648
income06$6 000 TO 6 999	0.8787	0.7196	1.2211	0.2220
income06$7 000 TO 7 999	-0.3046	0.7082	-0.4301	0.6671
income06$8 000 TO 9 999	0.1813	0.6661	0.2722	0.7855
income06$10000 TO 12499	-0.1764	0.5749	-0.3069	0.7589
income06$12500 TO 14999	-0.1069	0.5972	-0.1790	0.8579
income06$15000 TO 17499	0.0132	0.5959	0.0222	0.9823
income06$17500 TO 19999	0.3846	0.6134	0.6270	0.5306
income06$20000 TO 22499	-0.2002	0.5945	-0.3368	0.7363
income06$22500 TO 24999	-0.5127	0.6031	-0.8502	0.3952
income06$25000 TO 29999	-0.4333	0.5718	-0.7577	0.4486
income06$30000 TO 34999	-0.4376	0.5812	-0.7529	0.4515
income06$35000 TO 39999	0.0642	0.5722	0.1121	0.9107
income06$40000 TO 49999	-0.5743	0.5515	-1.0414	0.2977
income06$50000 TO 59999	-0.2330	0.5601	-0.4159	0.6775
income06$60000 TO 74999	-0.3882	0.5530	-0.7020	0.4827
income06$75000 TO $89999	0.1384	0.5568	0.2486	0.8037
income06$90000 TO $109999	-0.0203	0.5714	-0.0355	0.9717
income06$110000 TO $129999	-0.1165	0.6038	-0.1930	0.8470
income06$130000 TO $149999	-0.1718	0.6274	-0.2739	0.7842
income06$150000 OR OVER	-0.2112	0.5591	-0.3777	0.7056
owngunNO	-0.0575	0.1373	-0.4189	0.6753
partyid_3Ind	0.0972	0.1410	0.6896	0.4905
partyid_3Rep	0.0429	0.1975	0.2172	0.8281
polviewsLiberal	-0.0194	0.3080	-0.0629	0.9498
polviewsSlghtLib	0.4465	0.3219	1.3869	0.1655
polviewsModerate	-0.0194	0.2909	-0.0669	0.9467
polviewsSlghtCons	0.2840	0.3227	0.8801	0.3788
polviewsConserv	0.3146	0.3352	0.9387	0.3479
polviewsExtrmCons	1.4765	0.4494	3.2858	0.0010
prayONCE A DAY	0.1037	0.1521	0.6820	0.4953
praySEVERAL TIMES A WEEK	0.1817	0.2163	0.8403	0.4008
prayONCE A WEEK	0.2509	0.2469	1.0159	0.3097
prayLT ONCE A WEEK	-0.1799	0.2109	-0.8528	0.3938
prayNEVER	0.3063	0.2105	1.4548	0.1457
sexFemale	-0.0898	0.1207	-0.7439	0.4570
social_cons3Mod	0.2520	0.1632	1.5444	0.1225
social_cons3Conserv	0.3947	0.1908	2.0682	0.0386
southSouth	0.3533	0.1278	2.7634	0.0057
tolerance	0.6104	0.0335	18.2263	0.0000
wordsum	-0.0083	0.0377	-0.2203	0.8256
zodiacTAURUS	0.4438	0.2842	1.5619	0.1183
zodiacGEMINI	-0.0361	0.2783	-0.1296	0.8969
zodiacCANCER	-0.2522	0.2881	-0.8754	0.3814
zodiacLEO	0.2507	0.2686	0.9337	0.3505
zodiacVIRGO	-0.0382	0.2849	-0.1340	0.8934

Variable	Coefficient	Std. Error	T-Statistic	P Value
zodiacLIBRA	-0.0151	0.2735	-0.0553	0.9559
zodiacSCORPIO	0.1258	0.2883	0.4364	0.6625
zodiacSAGITTARIUS	0.3752	0.2922	1.2840	0.1991
zodiacCAPRICORN	-0.0918	0.2903	-0.3164	0.7517
zodiacAQUARIUS	0.2408	0.2784	0.8649	0.3871
zodiacPISCES	0.4617	0.2899	1.5929	0.1112

The aic values for the first model is 2056.9278067. If we evaluate the model by calculating how many prediction errors the model made, the error rate for model 1 is 0.2362609.

LASSO Regression (Model 2)

Besides just simply relying on evaluating how many prediction errors the model made to do model selection, the Lasso regression model offers another approach when we want to choose the fit model to predict future outcomes. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Lasso selection allows you to regularize ("shrink") coefficients. This means that the estimated coefficients are pushed towards 0, to make them work better on new data-sets ("optimized for prediction"). This allows you to use complex models and avoid over-fitting at the same time.

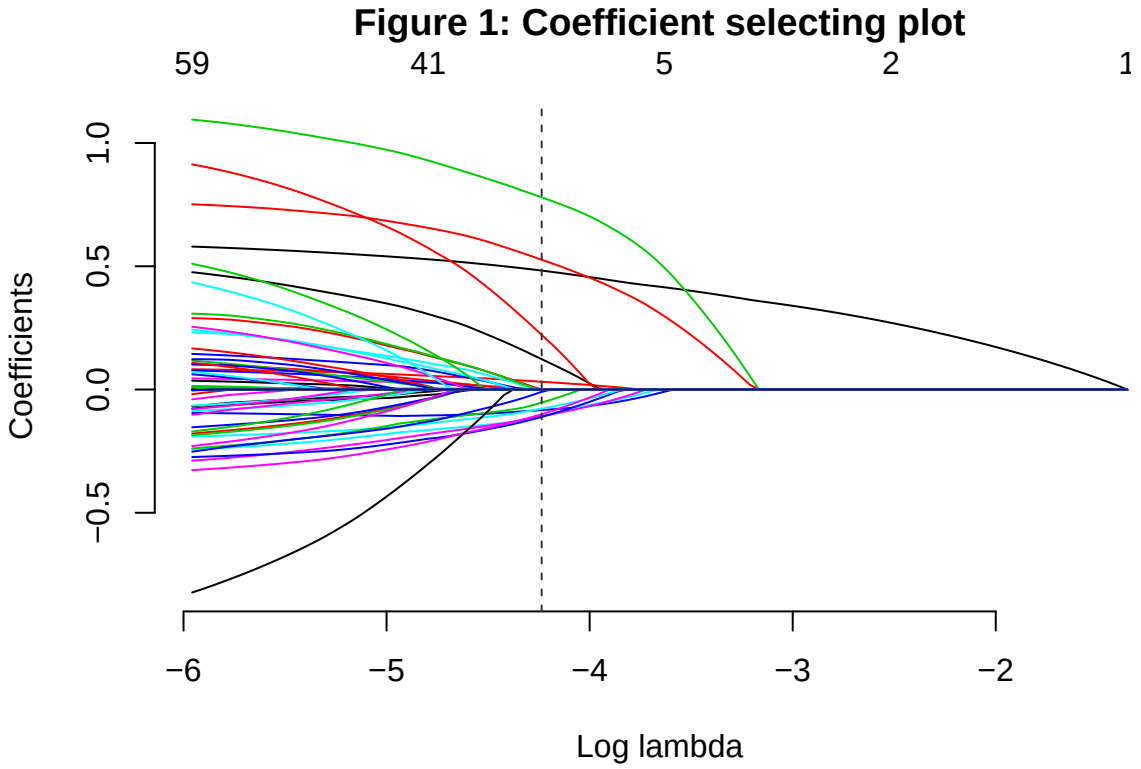

Figure 1 shows the coefficient selecting plot by lasso regression. The dotted line in the middle is the cut line where some coefficients are dropped (or shrink to 0), and some are kept with the optimal lambda value.

Here is the result of selected variables with their coefficients by lasso regression.

```
## 79 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                                 seg63
## intercept                  -5.293060143
## age                         0.002225277
## authoritarianism            0.031439945
## blackNo                     .
## blackYes                    .
## degree<HS                   0.526378291
## degreeHS                    .
## degreeCollege               .
## degreeGraduate deg          .
## egalit_scale                .
## grassLEGAL                  .
## grassNOT LEGAL              .
## hispanic_2No                .
## hispanic_2Yes               .
## income06UNDER $1 000        .
## income06$1 000 TO 2 999     .
## income06$3 000 TO 3 999     .
## income06$4 000 TO 4 999     .
## income06$5 000 TO 5 999     .
## income06$6 000 TO 6 999     0.221647263
## income06$7 000 TO 7 999     .
## income06$8 000 TO 9 999     .
## income06$10000 TO 12499     .
## income06$12500 TO 14999     .
## income06$15000 TO 17499     .
## income06$17500 TO 19999     0.123019556
## income06$20000 TO 22499     .
## income06$22500 TO 24999     .
## income06$25000 TO 29999     .
## income06$30000 TO 34999     .
## income06$35000 TO 39999     .
## income06$40000 TO 49999    -0.100323471
## income06$50000 TO 59999     .
## income06$60000 TO 74999     .
## income06$75000 TO $89999    .
## income06$90000 TO $109999   .
## income06$110000 TO $129999  .
## income06$130000 TO $149999  .
## income06$150000 OR OVER     .
## owngunYES                   .
## owngunNO                    .
## owngunREFUSED               .
## partyid_3Dem               -0.083537820
## partyid_3Ind                .
## partyid_3Rep                .
## polviewsExtrmLib            .
## polviewsLiberal            -0.053301625
## polviewsSlghtLib            .
## polviewsModerate           -0.108707934
## polviewsSlghtCons           .
## polviewsConserv             .
## polviewsExtrmCons           0.779826017
## praySEVERAL TIMES A DAY     .
```

```
## prayONCE A DAY              .
## praySEVERAL TIMES A WEEK    .
## prayONCE A WEEK             .
## prayLT ONCE A WEEK         -0.111159013
## prayNEVER                   .
## sexMale                     .
## sexFemale                   .
## social_cons3Liberal        -0.077723961
## social_cons3Mod             .
## social_cons3Conserv         .
## southNonsouth              -0.108720755
## southSouth                  0.000141765
## tolerance                   0.482460560
## wordsum                     .
## zodiacARIES                 .
## zodiacTAURUS                .
## zodiacGEMINI                .
## zodiacCANCER               -0.009532849
## zodiacLEO                   .
## zodiacVIRGO                 .
## zodiacLIBRA                 .
## zodiacSCORPIO               .
## zodiacSAGITTARIUS           .
## zodiacCAPRICORN             .
## zodiacAQUARIUS              .
## zodiacPISCES                .
```

The second model select variables based on the result from lasso regression. The second model selects relatively fewer variables than the first model. It ignores whether repspondent is African American, gebder, egalitarianism scale, attitude on legal marijuana, hispanic race, party identification, wordsum (number words correct in vocab test) and zodiac. The regression model is:

- colrac = beta0 + beta1age + beta2degree + beta3owngun + beta4polviews + beta5social_cons3 + beta6south + beta7tolerance

So how does it perform? The aic for this model2 is 2011.661829, lower than the model 1's aic value.

What about the error rate for model 2?

It is actually 0.2470467, roughly 0.0107858 higher. Even though the results seem contradict to each other, I would consider model2 as the better model, because it might be possible that the reason model 1 has lower error rate than model 2 is because model 1 have all the variables included. So the rest of the analysis will focus on model 2.

Interpreting the coefficients

Interestingly, the model didn't select some variables which I initially thought were very important factors, such as race and sex. Since the number of coefficients is large, it might be difficult to present the result through coefficient plot. Instead, here is the table:

Table 2: Logistic Regression Results

Variable	Coefficient	Std. Error	T-Statistic	P Value
(Intercept)	-6.3577	0.5134	-12.3832	0.0000

Variable	Coefficient	Std. Error	T-Statistic	P Value
age	0.0063	0.0032	1.9283	0.0538
degreeHS	-0.9272	0.1744	-5.3165	0.0000
degreeCollege	-1.0230	0.1954	-5.2366	0.0000
degreeGraduate deg	-1.1130	0.2359	-4.7182	0.0000
owngunNO	-0.0656	0.1284	-0.5111	0.6093
polviewsLiberal	-0.0554	0.2997	-0.1847	0.8534
polviewsSlghtLib	0.4011	0.3119	1.2858	0.1985
polviewsModerate	-0.0091	0.2779	-0.0327	0.9739
polviewsSlghtCons	0.2604	0.3046	0.8548	0.3926
polviewsConserv	0.3328	0.3096	1.0746	0.2825
polviewsExtrmCons	1.5582	0.4232	3.6823	0.0002
social_cons3Mod	0.2719	0.1396	1.9478	0.0514
social_cons3Conserv	0.3994	0.1616	2.4708	0.0135
southSouth	0.3207	0.1196	2.6812	0.0073
tolerance	0.5993	0.0311	19.2999	0.0000

Checking from P-value, we found that degree, political identification as extreme conservative, being socially conservative, coming from South and tolerance are significant variables we should consider. In particular, there are several variables that influence the attitude at a higher degree:

- Compared with being a conservative, an extreme conservative people will increase the odd to allow racist to teach in college by 0.96 (1.558-0.333-0.260).
- Compared with a high school degree people, a graduate degree people will become less likely to allow racist to teach in college by 0.186.

Overall, we can see that the factors that affecting attitude on allowing racist to teach in college are related to:

- The level of people get educated
- The political identifications
- The attitude towards social conservatism
- Whether people is from South

Limitations

The analysis is still short in many aspects, especially on the build of model. The first is the problem of multicollinearity that there might be linear relationship between variables. Some of the variables might have joint effect that we didn't discover, making the model less powerful in prediction. Since most variables are categorical variables, it is hard to decide when and how to add polynomial terms in the model. In particular, the income in this data is categorized as factor, rather than numeric variables, so I cannot add squared term to the income. In addition, other confounding variables: there might be confounding variables that the model fail to include, and thus the result is not accurate and the model is not stable.

Reference

lasso: https://stats.stackexchange.com/questions/251708/when-to-use-ridge-regression-and-lasso-regression-what-can-be-achieved-while-us