Homework Challenge

刘晶芳 (15320171151900)　余星月 (15320171151888)

1. 数据来源

中国综合社会调查 2015（Chinese General Social Survey，CGSS 2015）

2. 变量选择

(1) 被解释变量：个体是否幸福

(2) 解释变量（55个）：

个人层面：性别、年龄、个人年收入、个人教育程度、政治面貌、身体健康、户口、参加宗教活动频率、心理健康状况、互联网使用频率、学习频率、休息频率、社交频率、度假频率、工作单位类型、工作是否全职、是否结婚/同居、社会经济地位等级、收入是否合理、工作待遇是否合理（责任承担）、工作待遇是否合理（业绩）、心情平静频率、感到有活力的频率、非常累的频率、感到不能忍受的频率、换工作个数、教育是否与工作匹配、技能是否与工作匹配、做家务时间、通勤时间;

家庭层面：家庭收入、家庭房产套数、是否有车、子女数、配偶年龄、配偶受教育程度、配偶政治面貌、配偶收入、配偶周工作时间、配偶工作性质、父亲受教育程度、母亲受教育程度、14岁时候父亲就业状况;

公共服务评价：公共教育满意度、公共医疗满意度、住房保障满意度、就业保障满意度、社会管理满意度、社会保障满意度、劳动就业满意度、基础设施满意度;

社会态度：社会公平度、自我评价阶层、自我预期;

政治参与：是否参与选举

3. 实证结果及分析

我们分别对55个变量进行了logit回归、lasso回归以及elastic net回归，并对三种方法结果进行比较，分析三种方法的优劣。

（1）　方法一：logit回归

Logistic regression

life_satisfied	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig

gender	-0.205	0.821	-0.25	0.803	-1.813	1.404	
age	0.146	0.087	1.68	0.092	-0.024	0.317	*
income	0.000	0.000	-0.20	0.845	0.000	0.000	
edu	-0.815	0.439	-1.86	0.063	-1.675	0.045	*
party	1.049	0.570	1.84	0.066	-0.068	2.166	*
health	0.943	0.466	2.03	0.043	0.031	1.856	**
hukou_residence	-0.082	0.485	-0.17	0.866	-1.032	0.868	
belief	1.478	0.833	1.77	0.076	-0.155	3.111	*
mental	0.746	0.420	1.78	0.076	-0.077	1.568	*
inter_using	-0.388	0.323	-1.20	0.229	-1.021	0.244	
learning	0.953	0.562	1.70	0.090	-0.148	2.055	*
relaxing	0.708	0.610	1.16	0.246	-0.488	1.904	
social	0.393	0.456	0.86	0.389	-0.501	1.287	
holiday	0.337	0.276	1.22	0.223	-0.205	0.879	
job_type	1.142	0.401	2.85	0.004	0.357	1.927	***
job_nature	4.275	1.252	3.42	0.001	1.821	6.728	***
o.married	0.000	.	.	.	.	.	
social_ecolevel	0.683	0.719	0.95	0.342	-0.726	2.091	
income_proper	1.433	0.810	1.77	0.077	-0.154	3.021	*
jobduty_treat	-0.083	0.158	-0.52	0.601	-0.392	0.227	
jobcomple_treat	0.350	0.178	1.97	0.049	0.001	0.699	**
peacemood_frequency	-1.252	0.558	-2.24	0.025	-2.346	-0.159	**
energetic_frequency	0.409	0.444	0.92	0.358	-0.462	1.280	
exhaust_frequency	0.322	0.277	1.16	0.245	-0.221	0.865	
cannot_bear_freque~y	0.108	0.456	0.24	0.812	-0.785	1.001	
jobnumber	0.088	0.171	0.51	0.609	-0.248	0.423	
edu_match	1.492	1.009	1.48	0.139	-0.485	3.469	

skill_match	-1.099	1.117	-0.98	0.325	-3.289	1.091	
usual_housework_time	-0.355	0.343	-1.03	0.301	-1.027	0.318	
week_housework_time	0.564	0.281	2.01	0.044	0.014	1.114	**
commutetime	0.522	0.511	1.02	0.306	-0.478	1.523	
hh_totalincome	0.000	0.000	-1.92	0.054	0.000	0.000	*
hh_housenumber	2.167	0.919	2.36	0.018	0.365	3.969	**
hh_car	0.618	0.942	0.66	0.512	-1.228	2.464	
o.childnumber	0.000	.	.	.	.	.	
spouseage	-0.229	0.105	-2.19	0.028	-0.434	-0.024	**
spouseedu	0.837	0.377	2.22	0.026	0.099	1.575	**
spouseparty	-2.255	1.458	-1.55	0.122	-5.113	0.602	
spouseincome	0.000	0.000	1.01	0.313	0.000	0.000	
spouse_weekhour	-0.091	0.033	-2.79	0.005	-0.155	-0.027	***
spousejob_nature	-1.014	1.571	-0.65	0.519	-4.094	2.066	
fatheredu	0.043	0.394	0.11	0.912	-0.729	0.816	
motheredu	-0.036	0.405	-0.09	0.929	-0.831	0.759	
father_job14	-1.762	1.474	-1.20	0.232	-4.650	1.126	
comedu_degree	0.008	0.040	0.20	0.838	-0.070	0.087	
commedi_degree	-0.007	0.027	-0.25	0.804	-0.059	0.046	
housesupport_degree	-0.028	0.030	-0.93	0.352	-0.086	0.031	
commanage_degree	-0.041	0.041	-1.01	0.312	-0.121	0.039	

laboremploy_degree	-0.029	0.030	-0.95	0.343	-0.088	0.031	
sociasupport_degree	0.035	0.032	1.10	0.273	-0.028	0.098	
basic_manufac_degree	-0.006	0.022	-0.27	0.786	-0.049	0.037	
social_justice	1.301	0.444	2.93	0.003	0.431	2.172	***
self_level	-0.268	0.271	-0.99	0.323	-0.799	0.263	
forward_selflevel	0.114	0.096	1.19	0.233	-0.074	0.302	
politic_join	-1.909	0.970	-1.97	0.049	-3.809	-0.009	**
Constant	-10.397	5.094	-2.04	0.041	-20.381	-0.414	**

Mean dependent var	0.826	SD dependent var	0.380
Pseudo r-squared	0.553	Number of obs	258.000
Chi-square	132.169	Prob > chi2	0.000
Akaike crit. (AIC)	214.648	Bayesian crit. (BIC)	406.508

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

将54个变量直接进行logit回归分析，我们得到了25个变量呈显著。接下来我们进行lasso回归。

（2） 方法二：Lasso 回归

Lasso 估计量求解以下最小化问题：

$$\hat{\beta}_{lasso} = \arg\min_{\beta}(Y - X\beta)^T(Y - X\beta) + \lambda\left\|\beta\right\|_1$$

其中，λ 为微调参数，控制惩罚的力度而$\left\|\beta\right\|_1$ 为各回归系数的绝对值之和，lasso 为收缩估计量，即相比较 OLS 估计量，更为原点收缩。

Lasso

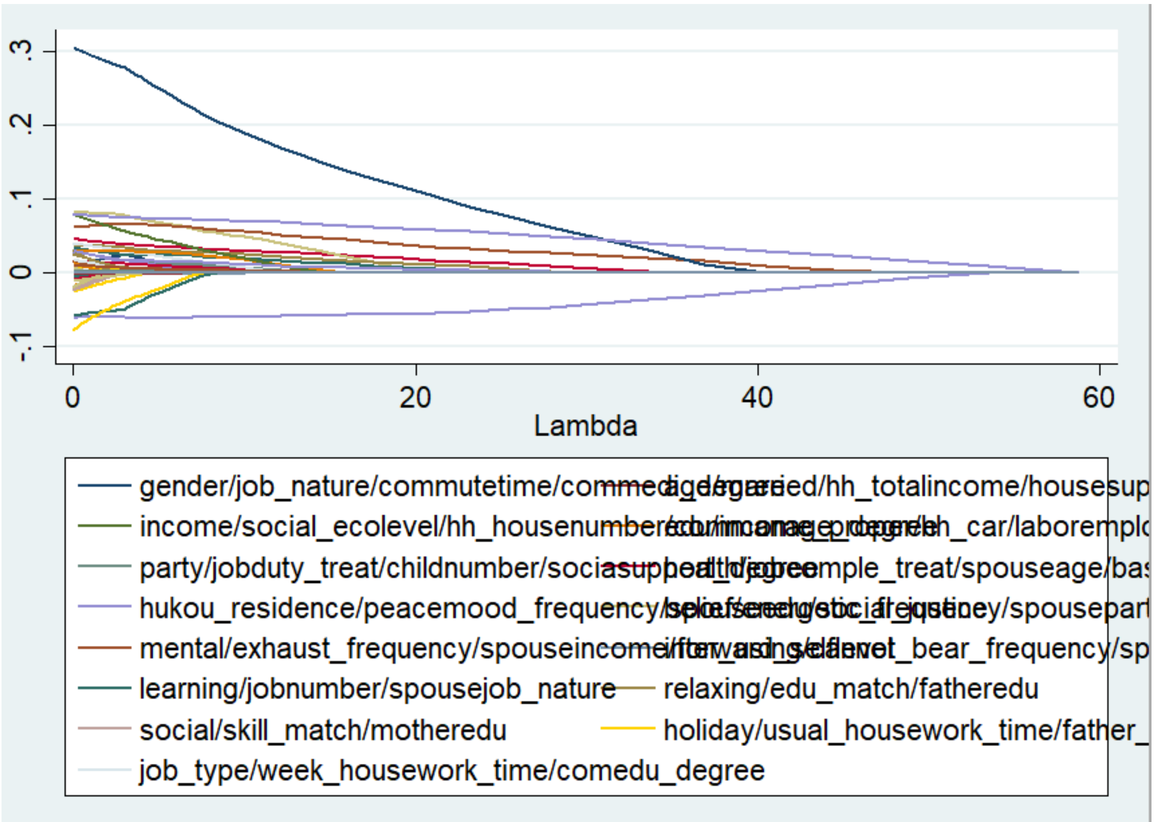
Knot|	ID	Lambda	s	L1-Norm	EBIC	R-sq
1|	1.000	58.700	1.000	0.000	-491.980	0.000
2|	2.000	53.485	2.000	0.009	-487.933	0.015
3|	3.000	48.734	3.000	0.024	-485.207	0.035
4|	4.000	44.404	4.000	0.043	-483.432	0.058
5|	6.000	36.865	5.000	0.092	-487.191	0.100
6|	7.000	33.590	6.000	0.128	-485.100	0.121
7|	9.000	27.887	8.000	0.195	-479.556	0.156
8|	10.000	25.410	9.000	0.226	-476.789	0.173

9|	11.000	23.153	10.000	0.256	-473.582	0.188
10|	13.000	19.222	11.000	0.316	-473.846	0.214
11|	16.000	14.540	14.000	0.416	-461.070	0.247
12|	17.000	13.249	16.000	0.451	-448.619	0.258
13|	18.000	12.072	18.000	0.487	-436.124	0.268
14|	21.000	9.132	19.000	0.593	-438.096	0.295
15|	22.000	8.321	20.000	0.622	-432.536	0.302
16|	23.000	7.581	22.000	0.655	-418.623	0.308
17|	24.000	6.908	25.000	0.696	-396.880	0.314
18|	25.000	6.294	27.000	0.736	-382.859	0.319
19|	27.000	5.226	28.000	0.808	-377.983	0.328
20|	28.000	4.761	29.000	0.840	-371.228	0.331
21|	29.000	4.338	31.000	0.870	-356.294	0.334
22|	30.000	3.953	32.000	0.902	-349.542	0.337
23|	31.000	3.602	33.000	0.933	-342.651	0.340
24|	34.000	2.725	35.000	1.008	-329.259	0.347
25|	35.000	2.483	39.000	1.033	-298.453	0.350
26|	36.000	2.262	40.000	1.057	-291.468	0.352
27|	37.000	2.061	41.000	1.080	-284.306	0.354
28|	38.000	1.878	42.000	1.104	-277.016	0.356
29|	39.000	1.711	43.000	1.127	-269.602	0.358
30|	41.000	1.421	44.000	1.169	-262.523	0.360
31|	46.000	0.892	45.000	1.247	-255.704	0.363
32|	54.000	0.424	46.000	1.324	-248.181	0.364
33|	57.000	0.321	47.000	1.344	-240.240	0.364
34|	58.000	0.292	48.000	1.350	-232.242	0.364
35|	60.000	0.243	49.000	1.360	-224.263	0.365
36|	71.000	0.087	50.000	1.392	-216.283	0.365
Use	'long'	option	for	full	output.	Type

（由于页面限制，结果分两段显示）

added/removed				
Added	cons.			
Added	social justice.			
Added	peacemood frequency.			
Added	mental.			
Added	job nature.			
Added	health.			
Added	relaxing	spouseedu.		
Added	spouse weekhour.			
Added	learning.			
Added	belief.			
Added	income proper	housesupport degree	forward selflevel.	
Added	party	hh housenumber.		
Added	job type	jobcomple treat.		
Added	exhaust frequency.			
Added	age.			
Added	hukou residence	spousejob nature.		
Added	gender	father job14	sociasupport degree.	
Added	jobnumber	fatheredu.		
Added	social.			
Added	week housework time.			
Added	cannot bear frequency	usual housework time.		
Added	holiday	self level.	Removed	cannot bear frequency.
Added	inter using.			
Added	spouseage	comedu degree.		
Added	edu	edu match	motheredu	basic manufac degree.
Added	commutetime.			
Added	skill match.			
Added	hh car.			
Added	spouseparty.			
Added	commanage degree.			
Added	energetic frequency.			
Added	cannot bear frequency.			
Added	jobduty treat.			
Added	politic join.			
Added	commedi degree.			
Added	social ecolevel.			
e.g.	'lasso2,	lic(ebic)'	to	run

上表显示随着调整参数λ由大变小，越来越多的变量进入模型，比如λ=58.700，常数项首先进入模型，λ=53.485，social_justice 进入模型，以此类推。



上图为整个解的路径(作为λ的函数)，画出了不同变量回归系数的变化过程。其中，当λ=0 时（图上最左边），不存在惩罚项，故此是 Lasso 等价于 OLS，而当λ很大时，由于惩罚力度过大，所有变量系数均为 0。

使用不同的微调参数，可得到不同的 lasso 估计值，一般，选择使得此模型预测能力最强的参数，即 K 折交叉验证

将样本数据随机分为 10 个等分。将第 1 个子样本作为“验证集”（validation set）而保留不用，而使用其余 9 个子样本作为“训练集”（training set）来估计此模型，再以此预测第 1 个子样本，并计算第 1 个子样本的“均方预测误差”（Mean Squared Prediction Error）。

其次，将第 2 个子样本作为验证集，而使用其余 9 个子样本作为训练集来预测第 2 个子样本，并计算第 2 个子样本的 MSPE。以此类推，将所有子样本的 MSPE 加总，即可得整个样本的 MSPE。

最后，选择微调参数 ，使得整个样本的 MSPE 最小，故具有最佳的预测能力。

K 折交叉验证的结果如下:

K-fold cross-validation with 10 folds. Elastic net with alpha=1.

Fold 1 2 3 4 5 6 7 8 9 10

	Lambda	MSPE	st. dev.	
1	58.700	0.145	0.019	
2	53.485	0.144	0.019	
3	48.734	0.143	0.019	
4	44.404	0.141	0.019	
5	40.460	0.139	0.018	
6	36.865	0.137	0.017	
7	33.590	0.134	0.017	
8	30.606	0.132	0.016	
9	27.887	0.130	0.015	^
10	25.410	0.128	0.015	
11	23.153	0.126	0.014	
12	21.096	0.124	0.014	
13	19.222	0.123	0.013	
14	17.514	0.122	0.013	
15	15.958	0.121	0.013	
16	14.540	0.120	0.012	
17	13.249	0.120	0.012	
18	12.072	0.119	0.012	
19	10.999	0.119	0.012	
20	10.022	0.119	0.012	
21	9.132	0.118	0.012	*
22	8.321	0.118	0.011	

23	7.581	0.119	0.011
24	6.908	0.119	0.011
25	6.294	0.119	0.011
26	5.735	0.120	0.011
27	5.226	0.121	0.011
28	4.761	0.122	0.011
29	4.338	0.123	0.011
30	3.953	0.124	0.011
31	3.602	0.125	0.011
32	3.282	0.127	0.011
33	2.990	0.128	0.012
34	2.725	0.129	0.012
35	2.483	0.131	0.012
36	2.262	0.132	0.012
37	2.061	0.133	0.012
38	1.878	0.134	0.012
39	1.711	0.135	0.012
40	1.559	0.136	0.012
41	1.421	0.136	0.013
42	1.294	0.137	0.013
43	1.179	0.138	0.013
44	1.075	0.139	0.013
45	0.979	0.139	0.013

46	0.892	0.140	0.013
47	0.813	0.140	0.014
48	0.741	0.141	0.014
49	0.675	0.142	0.014
50	0.615	0.142	0.014
51	0.560	0.143	0.014
52	0.511	0.143	0.014
53	0.465	0.143	0.014
54	0.424	0.144	0.014
55	0.386	0.144	0.014
56	0.352	0.145	0.014
57	0.321	0.145	0.014
58	0.292	0.145	0.014
59	0.266	0.145	0.014
60	0.243	0.146	0.014
61	0.221	0.146	0.014
62	0.201	0.146	0.014
63	0.183	0.146	0.014
64	0.167	0.146	0.014
65	0.152	0.147	0.014
66	0.139	0.147	0.014
67	0.126	0.147	0.014
68	0.115	0.147	0.014

69	0.105	0.147	0.014
70	0.096	0.147	0.014
71	0.087	0.147	0.014
72	0.079	0.147	0.014
73	0.072	0.147	0.014
74	0.066	0.147	0.015
75	0.060	0.148	0.015
76	0.055	0.148	0.015
77	0.050	0.148	0.015
78	0.045	0.148	0.015
79	0.041	0.148	0.015
80	0.038	0.148	0.015
81	0.034	0.148	0.015
82	0.031	0.148	0.015
83	0.029	0.148	0.015
84	0.026	0.148	0.015
85	0.024	0.148	0.015
86	0.022	0.148	0.015
87	0.020	0.148	0.015
88	0.018	0.148	0.015
89	0.016	0.148	0.015
90	0.015	0.148	0.015
91	0.014	0.148	0.015

92	0.012	0.148	0.015
93	0.011	0.148	0.015
94	0.010	0.148	0.015
95	0.009	0.148	0.015
96	0.009	0.148	0.015
97	0.008	0.148	0.015
98	0.007	0.148	0.015
99	0.006	0.148	0.015
100	0.006	0.148	0.015

打星号处 λ=9.132，即为可使 MSPE 最小化的 值。与此最优 对应的 Lasso 估计结果为:

* lopt = the lambda that minimizes MSPE.
   Run model: cvlasso, lopt
^ lse = largest lambda for which MSPE is within one standard error of the minimal MSPE.
   Run model: cvlasso, lse
Estimate lasso with lambda=9.132 (lopt).

	CV	
Selected	Lasso	Post-est OLS
party	0.007	0.022
health	0.030	0.045
belief	0.051	0.087
mental	0.057	0.071
learning	0.019	0.030
relaxing	0.025	0.036
job_type	0.010	0.039
job_nature	0.198	0.272
income_proper	0.019	0.033
jobcomple_treat	0.004	0.015
peacemood_frequ~y	-0.061	-0.065
exhaust_frequency	0.001	0.008
hh_housenumber	0.024	0.048
spouseedu	0.011	0.012
spouse_weekhour	-0.002	-0.003
housesupport_de~e	-0.001	-0.002
social_justice	0.070	0.076
forward_selflevel	0.002	0.005
Partialled-out*		

_cons	0.012	-0.352

上表右边第一列即为 Lasso 所估计的变量系数。其中，除常数项外，只有 18 个变量的系数非零，而其余变量的系数为 0。考虑到作为收缩估计量的 Lasso 存在 bias，上表右边第二列汇报了“Post lasso”估计量的结果，即仅适用 lasso 进行变量筛选，然后扔掉 lasso 的回归系数，再对筛选出来的变量进行 OLS 回归。

（3）方法三：Elastic Net 回归

Knot|	ID	Lambda	s	L1-Norm	EBIC	R-sq
1|	1.000	587.001	1.000	0.000	-491.980	0.000
2|	2.000	534.853	2.000	0.004	-489.925	0.008
3|	3.000	487.338	3.000	0.013	-488.497	0.020
4|	4.000	444.045	4.000	0.026	-488.172	0.036
5|	6.000	368.654	5.000	0.061	-491.240	0.070
6|	7.000	335.903	6.000	0.088	-491.025	0.088
7|	8.000	306.063	7.000	0.115	-490.116	0.106
8|	9.000	278.873	8.000	0.143	-489.134	0.124
9|	10.000	254.099	10.000	0.170	-482.985	0.142
10|	11.000	231.525	11.000	0.198	-481.663	0.159
11|	12.000	210.957	12.000	0.225	-479.054	0.174
12|	13.000	192.216	14.000	0.252	-470.289	0.187
13|	16.000	145.405	15.000	0.349	-471.347	0.226
14|	17.000	132.487	17.000	0.381	-460.977	0.237
15|	18.000	120.717	19.000	0.416	-450.326	0.249
16|	19.000	109.993	20.000	0.451	-446.109	0.260
17|	20.000	100.222	21.000	0.486	-442.035	0.271
18|	22.000	83.206	23.000	0.549	-430.977	0.288
19|	23.000	75.814	24.000	0.578	-424.636	0.295
20|	24.000	69.079	28.000	0.615	-397.956	0.301
21|	25.000	62.942	31.000	0.655	-378.416	0.308
22|	26.000	57.351	30.000	0.695	-384.853	0.314
23|	27.000	52.256	31.000	0.732	-377.477	0.319
24|	29.000	43.384	33.000	0.799	-362.476	0.327
25|	30.000	39.530	34.000	0.832	-355.294	0.331
26|	32.000	32.818	35.000	0.895	-347.182	0.337
27|	33.000	29.903	36.000	0.923	-339.434	0.340
28|	34.000	27.246	38.000	0.950	-324.157	0.343
29|	35.000	24.826	41.000	0.977	-305.532	0.346
30|	36.000	22.620	43.000	1.005	-290.145	0.348
31|	39.000	17.111	43.000	1.086	-287.560	0.355
32|	40.000	15.591	44.000	1.109	-279.040	0.357
33|	49.000	6.749	45.000	1.258	-264.342	0.363
34|	53.000	4.652	46.000	1.300	-254.228	0.364

35|	56.000	3.519	47.000	1.323	-245.039	0.364
36|	58.000	2.922	48.000	1.338	-236.666	0.365
37|	62.000	2.014	49.000	1.360	-227.492	0.365
38|	73.000	0.724	50.000	1.391	-217.487	0.365
Use	'long'	option	for	full	output.	Type

（由于页面限制，结果分两段显示）

Entered/removed				
Added	cons.			
Added	social justice.			
Added	peacemood frequency.			
Added	mental.			
Added	job nature.			
Added	health.			
Added	relaxing.			
Added	spouseedu.			
Added	learning	spouse weekhour.		
Added	energetic frequency.			
Added	forward selflevel.			
Added	belief	income proper.		
Added	housesupport degree.			
Added	party	hh housenumber.		
Added	jobcomple treat	exhaust frequency.		
Added	job type.			
Added	cannot bear frequency.			
Added	age	jobnumber.		
Added	hukou residence.			
Added	gender	social	spousejob nature	father job14.
Added	week housework time	fatheredu	sociasupport degree.	
Removed	energetic frequency.			
Added	edu match.			
Added	holiday	self level.		
Added	usual housework time.			
Added	inter using.			

Added	basic manufac degree.			
Added	commutetime	comedu degree.		
Added	edu	spouseage	motheredu.	
Added	skill match	hh car.		
Added	commanage degree.	Removed	cannot bear frequency.	
Added	spouseparty.			
Added	energetic frequency.			
Added	politic join.			
Added	cannot bear frequency.			
Added	jobduty treat.			
Added	commedi degree.			
Added	social ecolevel.			
e.g.	'lasso2,	lic(ebic)'	to	run

这里，我们也进行了 Elastic Net 方法进行回归。Elastic Net 是 Ridge 和 Lasso 方法的结合，这里也有模型选择的功能。比较两者之间的回归结果我们可以看到，两种方法模型选择的变量大体相同，但存在着些许差异。首先，变量选择数量不同，这里，lasso 变量选择为 36，而 Elastic Net 变量选择为 38。其次，解释变量对被解释的影响程度也存在差异。