Hw4　Regularization

（刘娟伶，15320171151901）

数据来源：U.S. General Social Survey (GSS)—1972.dta

因变量：happy（幸福程度）

自变量：Marital. Race. Region. Sex. Wrkstat. Health. Educ. Relig. Satfin. Partyid. Childs. Incom16. Satjob. etc（包括婚姻状况、种族、地域、性别、工作状况、健康程度、教育程度、宗教信仰、政党、孩子数量、16 岁时家庭收入、工作满意程度等 57 个变量）。

summary 1y 57x

Variables	Obs	Mean	Std.Dev.	Min	Max	p1	p99	Skew.	Kurt.
happy	1606	1.862	.67	1	3	1	3	.167	2.205
adults	1609	2.234	.827	1	6	1	5	1.273	5.728
babies	1613	.382	.758	0	4	0	3	2.099	7.025
family16	1613	2.103	2.093	1	8	1	8	1.724	4.643
marital	1613	1.76	1.404	1	5	1	5	1.613	3.938
preteen	1613	.462	.874	0	6	0	4	2.184	8.103
race	1613	1.167	.379	1	3	1	2	1.923	5.149
reg16	1613	4.299	2.456	0	9	0	9	.334	2.163
region	1613	4.713	2.529	1	9	1	9	.385	1.878
sex	1613	1.5	.5	1	2	1	2	.001	1
teens	1612	.407	.775	0	6	0	3	2.094	7.638
wrkstat	1613	3.456	2.664	1	8	1	8	.377	1.349
health	1612	2.006	.844	1	4	1	4	.521	2.651
news	1611	1.606	1.078	1	5	1	5	1.846	5.492
res16	1610	3.465	1.606	1	6	1	6	.373	1.996
age	1608	44.951	17.102	18	89	19	82	.313	2.104
educ	1608	11.327	3.456	0	20	1	19	-.531	3.881
relig	1608	1.546	.909	1	5	1	5	1.967	6.562
satfin	1608	1.903	.737	1	3	1	3	.155	1.851
partyid	1607	2.506	2.191	0	7	0	7	.529	1.934
sibs	1606	3.761	2.369	0	7	0	7	.106	1.635
childs	1605	2.297	1.981	0	8	0	8	.847	3.358
class	1604	2.419	.647	1	4	1	4	-.168	2.677
attend	1600	4.331	2.585	0	8	0	8	-.21	1.654
finrela	1599	2.889	.746	1	5	1	5	-.18	3.441
incom16	1591	2.746	.783	1	5	1	5	-.237	3.482
finalter	1590	1.955	.904	1	3	1	3	.089	1.23
vote68	1588	1.417	.645	1	4	1	3	1.291	3.512
chldmore	1585	2.298	.994	1	4	1	4	.666	2.352
earnrs	1580	1.522	1.006	0	6	0	5	.923	4.543

fework	1577	1.346	.476	1	2	1	2	.646	1.418
racschol	1574	1.123	.328	1	2	1	2	2.301	6.295
gunlaw	1562	1.276	.447	1	2	1	2	1.003	2.005
mobile16	1562	1.871	.884	1	3	1	3	.255	1.328
chldidel	1552	3.202	1.699	0	8	0	8	1.431	4.884
abhlth	1539	1.131	.337	1	2	1	2	2.192	5.807
income72	1474	5.262	2.778	1	12	1	12	.382	2.496
fepres	1533	1.264	.441	1	2	1	2	1.073	2.152
abnomore	1528	1.603	.489	1	2	1	2	-.42	1.176
abpoor	1507	1.512	.5	1	2	1	2	-.049	1.002
realinc	1474	28388.6	20552.38	2707	109355	2707	109355	1.533	6.519
incdef	1468	5.499	1.33	2	8	3	8	-.284	2.447
wrkslf	1448	1.897	.304	1	2	1	2	-2.614	7.833
isco68	1447	5192.493	2984.386	210	9996	320	9990	.135	1.783
prestige	1447	38.485	13.534	12	82	14	72	.241	2.746
industry	1451	541.64	298.801	17	999	17	999	-.309	1.708
pawrkslf	1364	1.63	.483	1	2	1	2	-.537	1.289
paisco68	1347	6268.985	2703.693	110	9996	240	9990	-.461	2.32
papres16	1347	38.863	12.157	12	82	16	74	.352	3.929
racjob	1332	1.032	.177	1	2	1	2	5.292	29.01
racobjct	1320	2.833	.406	1	3	1	3	-2.35	7.898
racmar	1309	1.607	.489	1	2	1	2	-.44	1.193
racpres	1265	1.262	.44	1	2	1	2	1.08	2.166
wksub	1176	1.205	.404	1	2	1	2	1.462	3.137
spwrksta	1158	3.456	2.775	1	8	1	8	.406	1.297
spisco68	1027	5147.711	2903.363	110	9996	290	9990	.171	1.858
satjob	944	1.697	.804	1	4	1	4	1.005	3.419

1. 首先考虑 lasso 回归

Knot	ID	Lambda	s	L1-Norm	EBIC	R-sq	Entered/removed			
1	1	128.385	1	0	-241.219	0	Added	_cons.		
2	2	116.98	4	0.028	-223.601	0.022	Added	satfin	finalter	satjob.
3	6	80.63	5	0.193	-247.06	0.12	Added	health.		
4	10	55.575	6	0.346	-259.022	0.177	Added	abhlth.		
5	12	46.139	8	0.443	-250.675	0.198	Added	class	attend.	
6	13	42.04	9	0.497	-247.025	0.21	Added	mobile16.		
7	14	38.306	12	0.555	-227.233	0.221	Added	sex	relig	fework.
8	15	34.903	13	0.628	-224.073	0.234	Added	preteen.		
9	17	28.977	15	0.759	-216.068	0.255	Added	reg16	pawrkslf.	
10	18	26.403	16	0.833	-212.491	0.266	Added	incdef.		
11	20	21.92	18	0.962	-203.464	0.283	Added	wrkstat	prestige.	

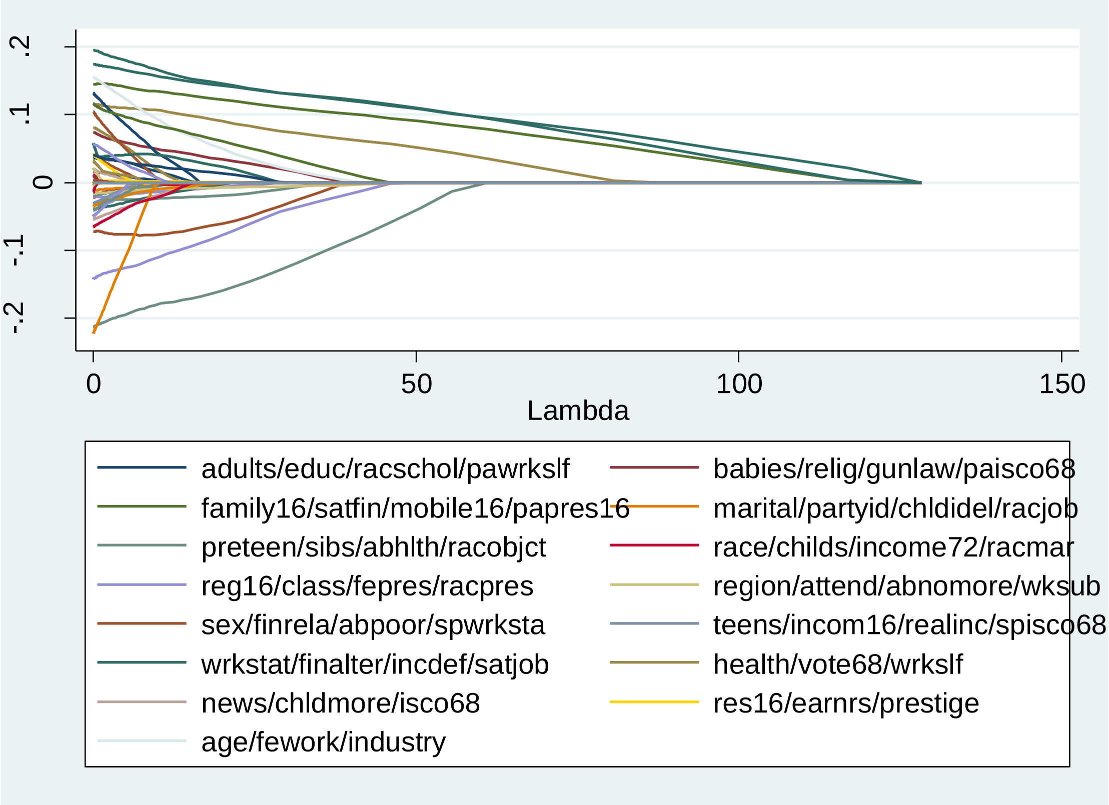
```
12   21   19.973   20   1.018    -190.394   0.29    Added     childs     chldidel.
13   23   16.582   21   1.117    -187.12    0.301   Added     news.
14   24   15.109   24   1.176    -165.205   0.307   Added     adults     racschol   racmar.
15   25   13.766   26   1.236    -151.26    0.312   Added     partyid    sibs.
16   26   12.543   28   1.298    -137.241   0.317   Added     educ       vote68.
17   27   11.429   30   1.356    -122.952   0.321   Added     teens      racpres.
18   28   10.414   31   1.417    -116.61    0.325   Added     babies.
19   29   9.489    33   1.49     -102.232   0.328   Added     abpoor     racjob.
20   30   8.646    32   1.575    -111.837   0.332   Removed   childs.

21   31   7.878    36   1.654    -81.002    0.335   Added     age        finrela    wrkslf   racobjct.
22   32   7.178    37   1.732    -74.178    0.338   Added     incom16.
23   34   5.959    39   1.874    -60.123    0.342   Added     chldmore   earnrs.
24   35   5.43     40   1.943    -52.95     0.344   Added     fepres.
25   41   3.107    41   2.263    -47.823    0.351   Added     spwrksta.
26   44   2.35     42   2.37     -40.297    0.352   Added     papres16.
27   46   1.951    43   2.429    -32.361    0.352   Added     gunlaw.
28   49   1.476    45   2.504    -16.363    0.352   Added     region     childs.
29   50   1.345    46   2.53     -8.352     0.352   Added     wksub.
30   57   0.701    47   2.667    -0.384     0.352   Added     income72.
31   58   0.639    48   2.682    7.814      0.352   Added     industry.
32   63   0.401    49   2.74     15.875     0.352   Added     abnomore.

33   65   0.333    50   2.758    23.942     0.352   Added     family16.
34   97   0.017    51   2.847    31.88      0.352   Added     res16.
Use 'long' option for full output. Type e.g. 'lasso2, lic(ebic)' to run the model selected by EBIC.
```

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=128.385 时，常数项首先进入模型。除常数项外，共有 33 个变量进入模型。

下图为整个解的路径（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（下图最左边），不存在惩罚项，故此时 Lasso 等价于 OLS。而当 λ 很大时（下图最右边），由于惩罚力度过大，所有变量系数均归于 0 。



接着，使用 Cross_Validation 的方法来选择最佳的调整参数，选择使 MSPE 最小的 λ，自行设定随机数种子，以便结果具有可重复性，默认 K=10。

	Lambda	MSPE	st.	dev.
1	128.385	0.434	0.028	
2	116.98	0.428	0.028	
3	106.588	0.417	0.028	
4	97.119	0.406	0.027	
5	88.491	0.397	0.027	
6	80.63	0.39	0.027	
7	73.467	0.383	0.027	
8	66.94	0.378	0.027	^
9	60.993	0.374	0.027	
10	55.575	0.37	0.027	
11	50.638	0.367	0.027	
12	46.139	0.365	0.027	
13	42.04	0.364	0.027	
14	38.306	0.362	0.027	
15	34.903	0.361	0.027	
16	31.802	0.359	0.027	
17	28.977	0.357	0.027	
18	26.403	0.355	0.027	
19	24.057	0.354	0.027	

20	21.92	0.353	0.026	*
21	19.973	0.353	0.026	
22	18.198	0.354	0.026	
23	16.582	0.355	0.026	
24	15.109	0.356	0.026	
25	13.766	0.357	0.025	
26	12.543	0.359	0.025	
27	11.429	0.36	0.025	
28	10.414	0.362	0.025	
29	9.489	0.364	0.025	
30	8.646	0.366	0.025	
31	7.878	0.368	0.025	
32	7.178	0.37	0.025	
33	6.54	0.373	0.025	
34	5.959	0.376	0.025	
35	5.43	0.378	0.025	
36	4.947	0.381	0.025	
37	4.508	0.383	0.025	
38	4.107	0.385	0.025	
39	3.743	0.387	0.026	
40	3.41	0.389	0.026	
41	3.107	0.391	0.026	
42	2.831	0.393	0.026	
43	2.58	0.394	0.026	
44	2.35	0.396	0.027	
45	2.142	0.397	0.027	
46	1.951	0.399	0.027	
47	1.778	0.401	0.027	
48	1.62	0.403	0.027	
49	1.476	0.404	0.028	
50	1.345	0.406	0.028	
51	1.226	0.407	0.028	
52	1.117	0.408	0.028	
53	1.017	0.409	0.028	
54	0.927	0.41	0.029	
55	0.845	0.411	0.029	
56	0.77	0.412	0.029	
57	0.701	0.413	0.029	

（因表格内容太多，已省略部分）

打星号处的 $\lambda=21.92$，是使 MSPE 最小的调整参数，即孩子的数量对该样本中幸福程度的影响最重要，与此对应的估计结果如下图所示：

Estimate lasso with lambda=21.92 (lopt).

Selected	Lasso	Post-est	OLS
preteen	-0.018	-0.039	
reg16	-0.005	-0.020	
sex	-0.057	-0.112	
wrkstat	-0.000	-0.029	
health	0.087	0.127	
relig	0.031	0.065	
satfin	0.120	0.143	
class	-0.070	-0.169	
attend	-0.007	-0.012	
finalter	0.141	0.168	
fework	0.042	0.123	
mobile16	0.056	0.102	
abhlth	-0.154	-0.229	
incdef	0.019	0.068	
prestige	0.000	0.003	
pawrkslf	0.011	0.070	
satjob	0.143	0.175	
Partialled-out*			
_cons	1.030	0.603	

上表“Lasso”所估计的变量系数中，除常数项外，只有 17 个变量的系数为非零，而其余变量（未出现在表中）的系数则为 0。考虑到作为收缩估计量的 Lasso 存在偏差（bias），上表“Post Lasso” 估计量的结果为，仅使用 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。

2. 再考虑 Elastic-net regression

下表是 alpha=0.2 的结果，R^2 仅为 0.3， 只有 18 个变量的系数不为零，结果不是很好。

```
Elastic-net regression                  Number of observations      =        298
                                             R-squared                    =     0.2967
                                             alpha                        =     0.2000
                                             lambda                       =     0.1424
                                             Cross-validation MSE     =     0.3592
                                             Number of folds              =         10
                                             Number of alpha tested     =          6
                                             Number of lambda tested    =        100


 happy                                  Coef.
```

adults	0
babies	0
family16	0
marital	0
preteen	-0.021
race	0
reg16	-0.007
region	0
sex	-0.058
teens	0
wrkstat	-0.008
health	0.091
news	0
res16	0
age	0
educ	0
relig	0.036
satfin	0.119
partyid	0
sibs	0
childs	-0.003
class	-0.080
attend	-0.009
finrela	0
incom16	0
finalter	0.135
vote68	0
chldmore	0
earnrs	0
fework	0.057
racschol	0
gunlaw	0
mobile16	0.060
chldidel	0
abhlth	-0.154
income72	0
fepres	0
abnomore	0
abpoor	0
realinc	0
incdef	0.022
wrkslf	0
isco68	0
prestige	0.001

industry	0	
pawrkslf		0.020
paisco68		0.000
papres16	0	
racjob	0	
racobjct	0	
racmar	0	
racpres	0	
wksub	0	
spwrksta	0	
spisco68	0	
satjob		0.137
_cons		0.962