REPORT

杨帆 2020103661 流行病与卫生统计学

目录

1 正态性检验 **2**

1.1 box-cox 数据转换 3

1.2 归一化处理 3

2 Supervised screening **3**

3 拟合 logistic 回归，使用 lasso、MCP、L_0 penalty 进行变量选择 **11**

3.1 lasso 11

3.2 MCP 13

3.3 L_0 penalization 15

4 K means 聚类 **17**

5 Composite penalization **19**

5.1 group lasso 19

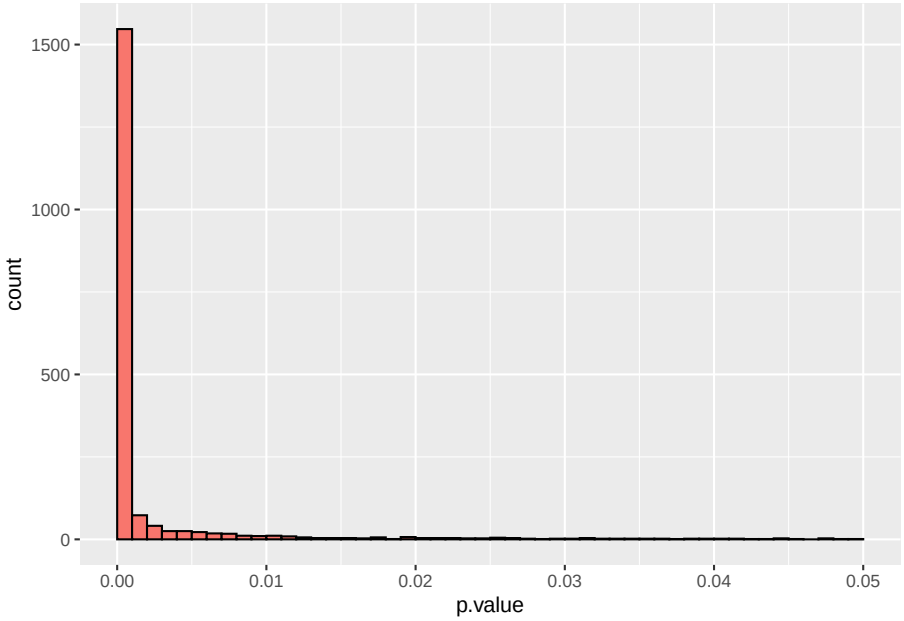
5.2 group MCP 21

6 lasso、MCP、L_0 penalt、group lasso 和 group MCP 结果总结 **23**

- 模型专业解释仅仅依据模型结果，不作为正式报告。
- 基因名按照顺序命名：Has1:Has2000。

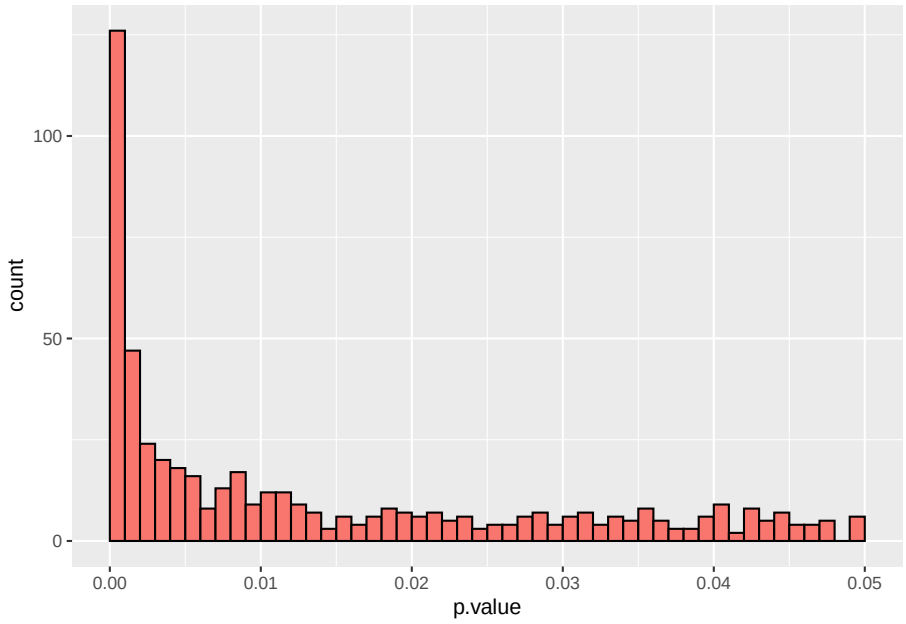
---

1 正态性检验



有 1912 个变量在 $\alpha = 0.05$ 水准下拒绝原假设，可认为不服从正态分布（尽管存在大量的假阳性）

1.1 box-cox 数据转换



经过 Box-cox 转换，依然有 527 个基因不服从正态分布。

故这里使用原数据采用 Wilcoxon 检验 2000 个基因在正常组织和肿瘤组织的分布是否相同。

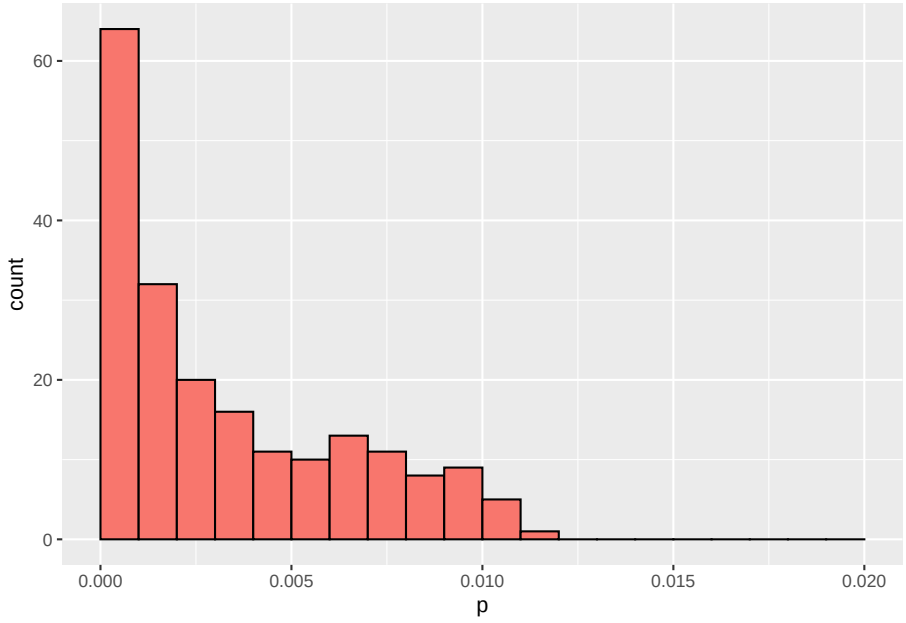
1.2 归一化处理

2 Supervised screening

采用 Wilcoxon 检验比较正常组织和肿瘤组织基因表达均值的差异

top 200 基因 p 值的分布

```
## Warning: `show.legend` must be a logical vector.
```



top 200 基因 wilcoxon 检验结果如下

.y.	group1	group2	n1	n2	statistic	p
Has493	0	1	22	40	778	0.000
Has1772	0	1	22	40	110	0.000
Has513	0	1	22	40	119	0.000
Has1042	0	1	22	40	119	0.000
Has1671	0	1	22	40	129	0.000
Has780	0	1	22	40	140	0.000
Has1582	0	1	22	40	145	0.000
Has625	0	1	22	40	150	0.000
Has377	0	1	22	40	729	0.000
Has1423	0	1	22	40	727	0.000
Has1060	0	1	22	40	155	0.000
Has897	0	1	22	40	723	0.000
Has249	0	1	22	40	722	0.000
Has1771	0	1	22	40	147	0.000
Has765	0	1	22	40	716	0.000

.y.	group1	group2	n1	n2	statistic	p
Has1635	0	1	22	40	715	0.000
Has964	0	1	22	40	166	0.000
Has365	0	1	22	40	167	0.000
Has245	0	1	22	40	710	0.000
Has1325	0	1	22	40	171	0.000
Has138	0	1	22	40	174	0.000
Has399	0	1	22	40	174	0.000
Has1153	0	1	22	40	177	0.000
Has66	0	1	22	40	702	0.000
Has1002	0	1	22	40	182	0.000
Has1730	0	1	22	40	182	0.000
Has267	0	1	22	40	697	0.000
Has391	0	1	22	40	186	0.000
Has1843	0	1	22	40	693	0.000
Has75	0	1	22	40	189	0.000
Has822	0	1	22	40	690	0.000
Has1411	0	1	22	40	690	0.000
Has1414	0	1	22	40	192	0.000
Has992	0	1	22	40	193	0.000
Has1494	0	1	22	40	687	0.000
Has1900	0	1	22	40	193	0.000
Has43	0	1	22	40	194	0.000
Has241	0	1	22	40	195	0.000
Has1346	0	1	22	40	195	0.000
Has1648	0	1	22	40	195	0.000
Has31	0	1	22	40	196	0.000
Has571	0	1	22	40	199	0.000
Has1770	0	1	22	40	199	0.000
Has26	0	1	22	40	201	0.000
Has187	0	1	22	40	202	0.000
Has1406	0	1	22	40	203	0.000
Has1808	0	1	22	40	203	0.000

.y.	group1	group2	n1	n2	statistic	p
Has47	0	1	22	40	205	0.000
Has698	0	1	22	40	207	0.000
Has127	0	1	22	40	208	0.000
Has802	0	1	22	40	209	0.000
Has1293	0	1	22	40	210	0.001
Has62	0	1	22	40	211	0.001
Has467	0	1	22	40	211	0.001
Has1263	0	1	22	40	211	0.001
Has1634	0	1	22	40	211	0.001
Has1679	0	1	22	40	212	0.001
Has1870	0	1	22	40	214	0.001
Has1067	0	1	22	40	215	0.001
Has619	0	1	22	40	216	0.001
Has1473	0	1	22	40	216	0.001
Has286	0	1	22	40	662	0.001
Has739	0	1	22	40	662	0.001
Has1549	0	1	22	40	218	0.001
Has415	0	1	22	40	659	0.001
Has72	0	1	22	40	222	0.001
Has515	0	1	22	40	222	0.001
Has652	0	1	22	40	222	0.001
Has1227	0	1	22	40	223	0.001
Has1983	0	1	22	40	223	0.001
Has1993	0	1	22	40	223	0.001
Has143	0	1	22	40	656	0.001
Has495	0	1	22	40	224	0.001
Has581	0	1	22	40	224	0.001
Has1967	0	1	22	40	656	0.001
Has1058	0	1	22	40	655	0.001
Has1208	0	1	22	40	226	0.001
Has994	0	1	22	40	227	0.001
Has83	0	1	22	40	228	0.002

.y.	group1	group2	n1	n2	statistic	p
Has295	0	1	22	40	228	0.002
Has590	0	1	22	40	228	0.002
Has141	0	1	22	40	229	0.002
Has1867	0	1	22	40	229	0.002
Has258	0	1	22	40	230	0.002
Has281	0	1	22	40	230	0.002
Has1110	0	1	22	40	230	0.002
Has1372	0	1	22	40	230	0.002
Has520	0	1	22	40	231	0.002
Has824	0	1	22	40	649	0.002
Has111	0	1	22	40	648	0.002
Has1221	0	1	22	40	232	0.002
Has1387	0	1	22	40	648	0.002
Has1674	0	1	22	40	648	0.002
Has1972	0	1	22	40	232	0.002
Has343	0	1	22	40	233	0.002
Has1884	0	1	22	40	647	0.002
Has529	0	1	22	40	234	0.002
Has1047	0	1	22	40	234	0.002
Has1256	0	1	22	40	234	0.002
Has1187	0	1	22	40	235	0.002
Has1511	0	1	22	40	235	0.002
Has201	0	1	22	40	644	0.002
Has360	0	1	22	40	236	0.002
Has989	0	1	22	40	236	0.002
Has1974	0	1	22	40	644	0.002
Has147	0	1	22	40	237	0.002
Has1959	0	1	22	40	238	0.003
Has107	0	1	22	40	239	0.003
Has830	0	1	22	40	239	0.003
Has1466	0	1	22	40	239	0.003
Has1897	0	1	22	40	640	0.003

.y.	group1	group2	n1	n2	statistic	p
Has440	0	1	22	40	241	0.003
Has812	0	1	22	40	639	0.003
Has1260	0	1	22	40	241	0.003
Has1340	0	1	22	40	241	0.003
Has1546	0	1	22	40	241	0.003
Has67	0	1	22	40	638	0.003
Has444	0	1	22	40	242	0.003
Has601	0	1	22	40	242	0.003
Has779	0	1	22	40	242	0.003
Has489	0	1	22	40	243	0.003
Has1334	0	1	22	40	243	0.003
Has1675	0	1	22	40	243	0.003
Has1115	0	1	22	40	244	0.003
Has1668	0	1	22	40	636	0.003
Has1904	0	1	22	40	244	0.003
Has264	0	1	22	40	245	0.004
Has1935	0	1	22	40	245	0.004
Has102	0	1	22	40	246	0.004
Has1902	0	1	22	40	246	0.004
Has1912	0	1	22	40	246	0.004
Has1965	0	1	22	40	246	0.004
Has1196	0	1	22	40	247	0.004
Has1810	0	1	22	40	247	0.004
Has86	0	1	22	40	248	0.004
Has100	0	1	22	40	248	0.004
Has627	0	1	22	40	248	0.004
Has1442	0	1	22	40	248	0.004
Has1583	0	1	22	40	248	0.004
Has1942	0	1	22	40	248	0.004
Has163	0	1	22	40	249	0.004
Has199	0	1	22	40	249	0.004
Has1799	0	1	22	40	251	0.005

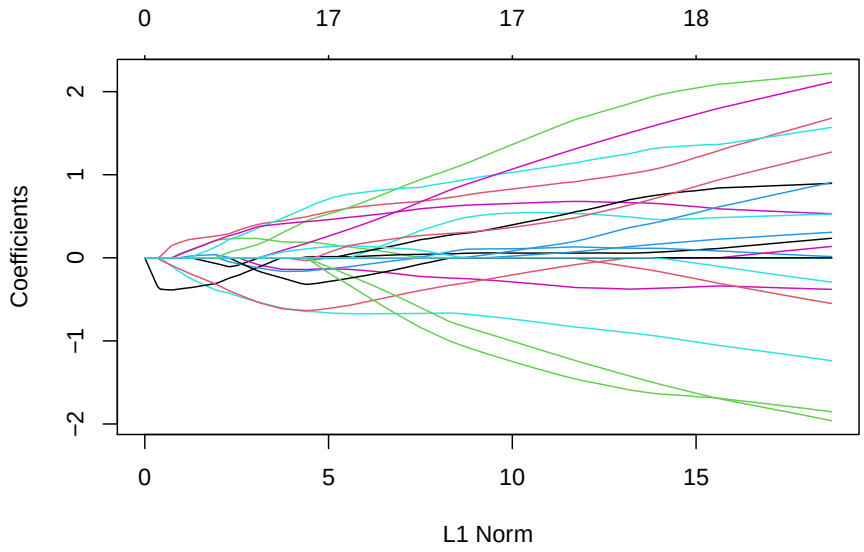
.y.	group1	group2	n1	n2	statistic	p
Has648	0	1	22	40	252	0.005
Has1248	0	1	22	40	252	0.005
Has1839	0	1	22	40	252	0.005
Has639	0	1	22	40	253	0.005
Has1111	0	1	22	40	627	0.005
Has15	0	1	22	40	254	0.006
Has1489	0	1	22	40	254	0.006
Has1637	0	1	22	40	254	0.006
Has1258	0	1	22	40	625	0.006
Has1836	0	1	22	40	625	0.006
Has91	0	1	22	40	256	0.006
Has437	0	1	22	40	624	0.006
Has550	0	1	22	40	257	0.007
Has661	0	1	22	40	257	0.007
Has807	0	1	22	40	623	0.007
Has1168	0	1	22	40	257	0.007
Has1194	0	1	22	40	257	0.007
Has1447	0	1	22	40	257	0.007
Has1886	0	1	22	40	257	0.007
Has1960	0	1	22	40	257	0.007
Has384	0	1	22	40	258	0.007
Has1366	0	1	22	40	258	0.007
Has1763	0	1	22	40	258	0.007
Has549	0	1	22	40	259	0.007
Has806	0	1	22	40	621	0.007
Has1247	0	1	22	40	621	0.007
Has1892	0	1	22	40	621	0.007
Has617	0	1	22	40	620	0.008
Has694	0	1	22	40	260	0.008
Has979	0	1	22	40	260	0.008
Has190	0	1	22	40	261	0.008
Has538	0	1	22	40	261	0.008

.y.	group1	group2	n1	n2	statistic	p
Has1353	0	1	22	40	261	0.008
Has1761	0	1	22	40	619	0.008
Has559	0	1	22	40	618	0.008
Has834	0	1	22	40	262	0.008
Has882	0	1	22	40	262	0.008
Has1887	0	1	22	40	262	0.008
Has14	0	1	22	40	617	0.009
Has85	0	1	22	40	263	0.009
Has662	0	1	22	40	617	0.009
Has1534	0	1	22	40	263	0.009
Has500	0	1	22	40	264	0.009
Has527	0	1	22	40	264	0.009
Has1472	0	1	22	40	264	0.009
Has958	0	1	22	40	265	0.009
Has1136	0	1	22	40	265	0.009
Has271	0	1	22	40	266	0.010
Has427	0	1	22	40	266	0.010
Has1073	0	1	22	40	266	0.010
Has1920	0	1	22	40	266	0.010
Has785	0	1	22	40	267	0.010
Has1413	0	1	22	40	267	0.010
Has1943	0	1	22	40	613	0.010
Has1659	0	1	22	40	268	0.011
Has1798	0	1	22	40	268	0.011
Has49	0	1	22	40	611	0.011

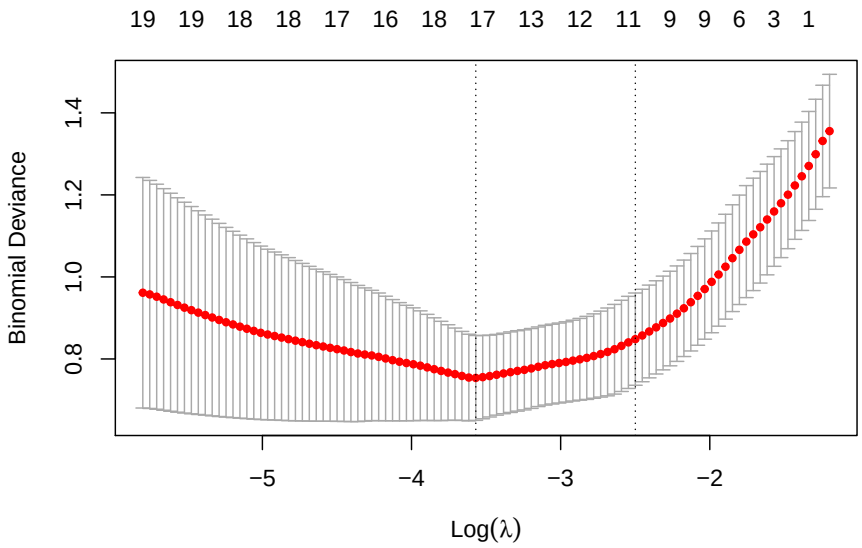
3 拟合 logistic 回归，使用 lasso、MCP、L_0 penalty 进行变量选择

3.1 lasso

lasso 解的路径



交叉验证结果



经过 5 折交叉验证，选取 $\lambda = 0.028$，拟合模型，选取基因如下：

gene_names	beta
Has1772	0.5847296
Has1582	0.1469483
Has377	-0.6703120
Has1423	-0.1333174
Has765	-0.5854061
Has1325	0.5767241
Has66	-0.1142634
Has1346	0.7446597
Has1870	0.4775175
Has1473	0.0096716
Has1221	0.3138660
Has1466	0.0274690
Has1668	-0.1742049
Has639	0.1414367
Has617	-0.2614381

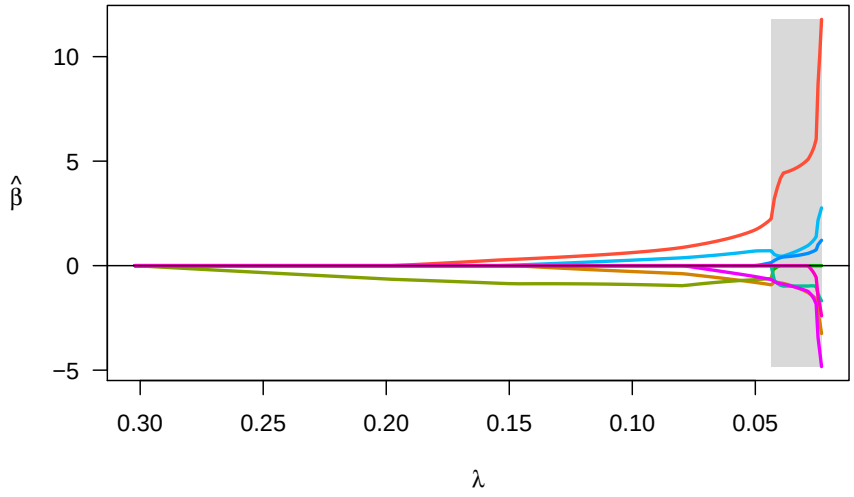
gene_names	beta
Has1353	0.1155853
Has14	-0.2786412

- Has1772、Has1582、Has1325、Has391、Has1346、Has1870、Has1993、Has590、Has1221、Has1546、Has627、Has527、Has271、Has427 回归系数为正数，表明这 14 个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。
- Has377、Has1423、Has765、Has66、Has286、Has1058、Has617、Has14 回归系数为负数，表明这 14 个基因在肿瘤组织表达更多，即其表达可能会抑制肿瘤的生长。

3.2 MCP

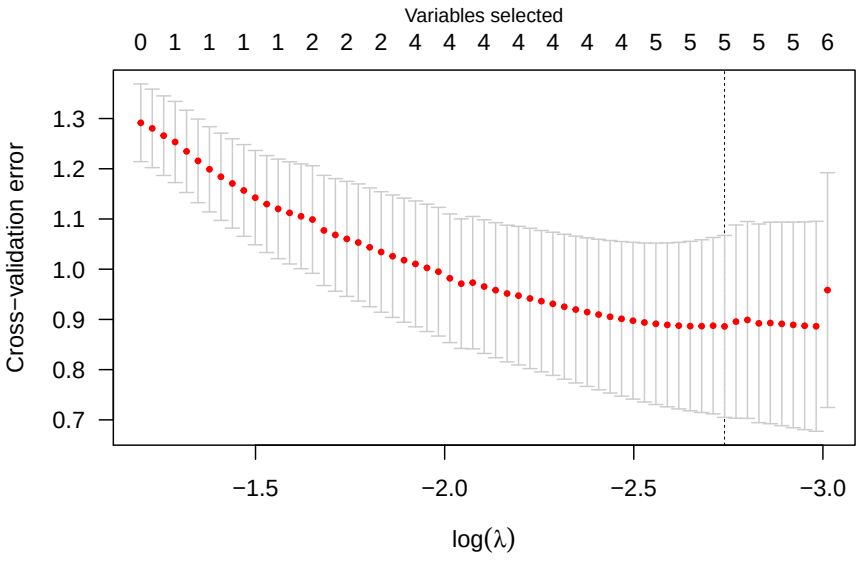
MCP 解的路径

```
## Warning in ncvreg(X, Y, family = "binomial"): Maximum number of iterations
## reached
```



交叉验证结果

```
## Warning in ncvreg(X = X, y = y, ...): Maximum number of iterations reached
```



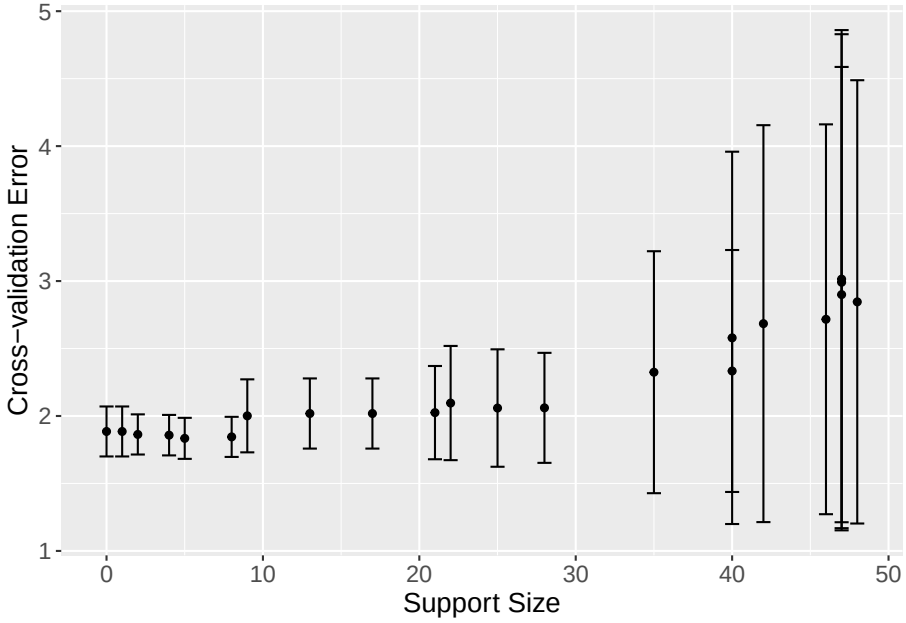
经过 5 折交叉验证，选取 $\lambda = 0.07$，拟合模型，选取基因如下：

gene_names	beta
Has1772	1.2002765
Has377	-0.5840983
Has249	-0.7870736
Has1870	0.5179080
Has617	-0.2597950

- Has1772、Has1870 回归系数是正数，表明这 2 个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。
- Has377、Has249、Has617 回归系数是负数，表明这 2 个基因在肿瘤组织表达更少，即其表达可能会抑制肿瘤的生长。

3.3 L_0 penalization

采用 AIC、BIC 逐步回归进行变量选择，均选择了 61 个基因，结果如下:



3 拟合 *LOGISTIC* 回归,使用 *LASSO*、*MCP*、*L_0 PENALTY* 进行变量选择16

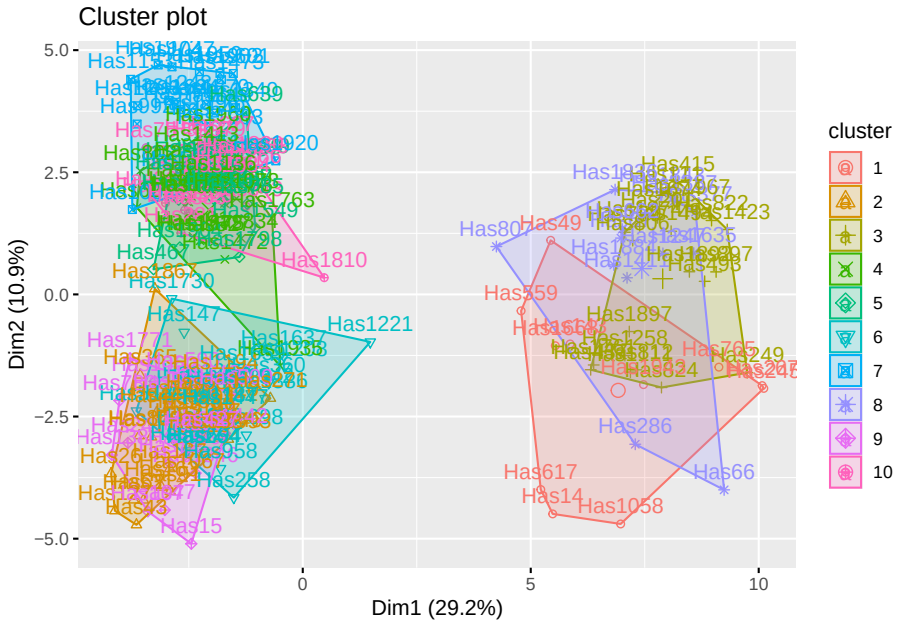
L_0 变量选择结果

gene_names	beta
Has377	-0.2172260
Has249	-0.2560676
Has1325	0.1041478
Has1466	0.1297356
Has806	0.1695493

- Has1325、Has1466、Has1546、Has1334、Has1248、Has806、Has527、Has1920 回归系数是正数，表明这 2 个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。
- Has377、Has249、Has495、Has779、Has86 回归系数是负数，表明这 2 个基因在肿瘤组织表达更少，即其表达可能会抑制肿瘤的生长。

4 K means 聚类

可视化结果，可以看到很多类有重合。



kmeans 聚类结果

gene_group	genes
1	Has765 , Has245 , Has267 , Has143 , Has1058, Has1668, Has617 , Has559 , Has14 , Has1943, Has49
2	Has1772, Has365 , Has43 , Has31 , Has571 , Has26 , Has187 , Has127 , Has62 , Has1067, Has619 , Has1227, Has1993, Has994 , Has141 , Has1867, Has1110, Has1546, Has444 , Has1334, Has86 , Has163 , Has199 , Has91 , Has550 , Has1194, Has538 , Has271
3	Has493 , Has1423, Has897 , Has249 , Has822 , Has1494, Has739 , Has415 , Has1967, Has824 , Has111 , Has1674, Has201 , Has1974, Has1897, Has812 , Has1111, Has1258, Has437 , Has806 , Has1892, Has1761
4	Has625 , Has391 , Has1406, Has802 , Has1679, Has1372, Has1511, Has830 , Has1260, Has1675, Has1115, Has1935, Has102 , Has1965, Has1583, Has1960, Has1763, Has834 , Has527 , Has1472, Has1136, Has1073, Has1413
5	Has1900, Has1770, Has467 , Has1549, Has495 , Has1256, Has989 , Has1904, Has100 , Has1799, Has639 , Has1168, Has190 , Has882 , Has1887, Has85 , Has785 , Has1798
6	Has513 , Has1730, Has1648, Has1808, Has698 , Has1263, Has652 , Has1208, Has258 , Has1221, Has1187, Has360 , Has147 , Has440 , Has264 , Has1912, Has1942, Has1489, Has1637, Has1447, Has1886, Has694 , Has958 , Has1659
7	Has1153, Has1002, Has992 , Has1293, Has1634, Has1870, Has1473, Has1983, Has581 , Has1972, Has1047, Has1959, Has1340, Has601 , Has779 , Has1902, Has1196, Has648 , Has1248, Has549 , Has1920
8	Has377 , Has1635, Has66 , Has1843, Has1411, Has286 , Has1387, Has1884, Has67 , Has1836, Has807 , Has1247, Has662
9	Has1042, Has1671, Has780 , Has1060, Has1771, Has138 , Has399 , Has241 , Has1346, Has47 , Has72 , Has515 , Has590 , Has107 , Has1466, Has1442, Has15 , Has384 , Has500 , Has427

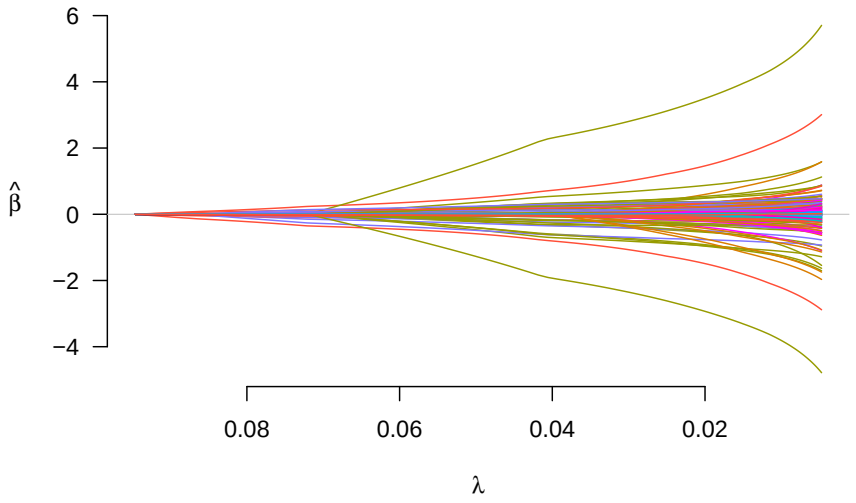
gene_genes	
10	Has1582, Has964 , Has1325, Has75 , Has1414, Has83 , Has295 , Has281 , Has520 , Has343 , Has529 , Has489 , Has1810, Has627 , Has1839, Has661 , Has1366, Has979 , Has1353, Has1534

5 Composite penalization

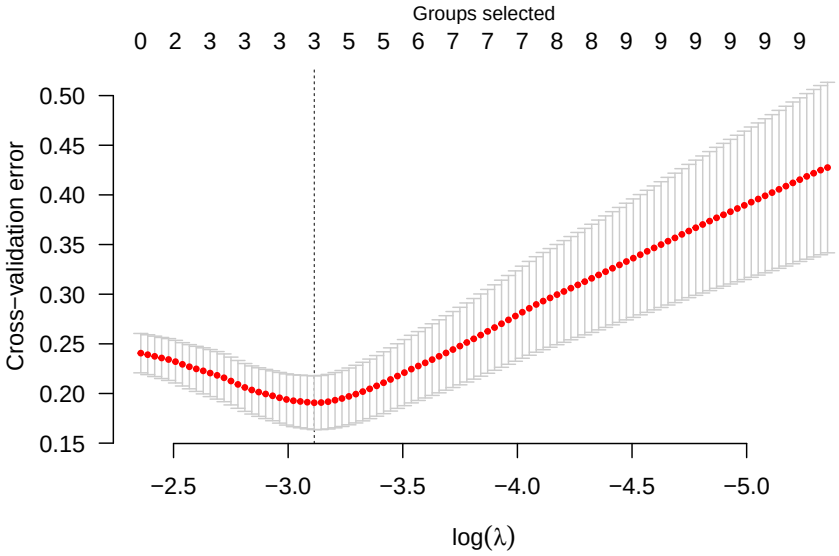
这里采用了两种方法，分别是 group lasso，group MCP。

5.1 group lasso

group lasso 解的路径



采用 5 折交叉验证的方法选择 λ.



group lasso 通过 5 折交叉验证选择 $\lambda = 0.044$，进行模型拟合，从 200 个基因中选出了 23 个变量, 分别为：

gene_names	beta
Has267	0.6092363
Has1411	0.1876647
Has1494	0.4618321
Has286	0.0020350
Has739	0.2321242
Has415	0.1389012
Has1058	0.0045176
Has824	1.9657277
Has111	0.2619316
Has1387	0.1891688
Has1674	0.1099478
Has1884	0.1217743
Has201	0.2878553
Has1974	0.2840208

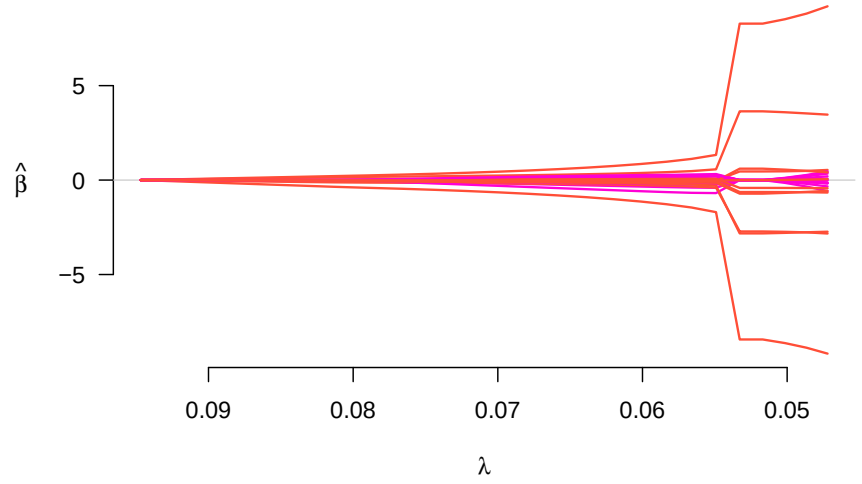
gene_names	beta
Has67	0.1848672
Has1111	0.2611245
Has807	0.0050405
Has1247	0.0648507
Has1892	0.0406205
Has1761	0.2142693
Has559	0.1724631
Has14	0.0134979
Has662	0.2773290

回归系数均为正数，表明这 23 个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。

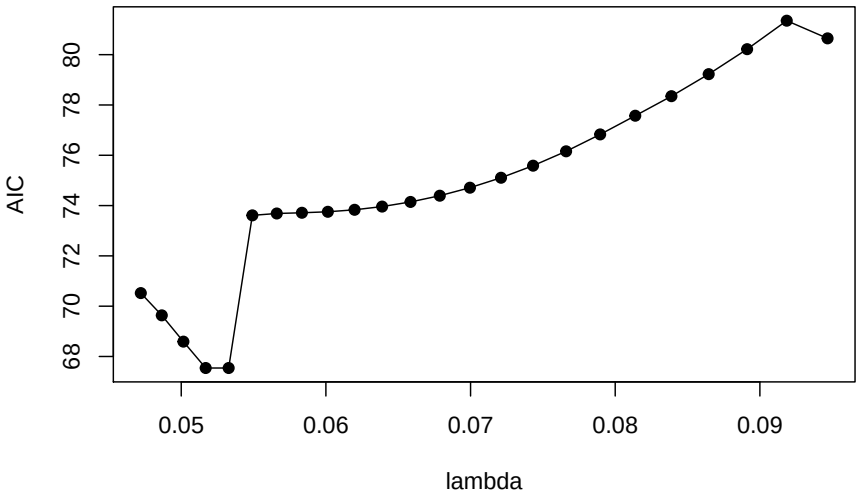
5.2 group MCP

group MCP 解的路径

```
## Warning in grpreg(X, Y, group, penalty = "grMCP", family = "binomial"): Model
## saturated; exiting...
```



使用 AIC 准则选取 λ，不同 λ 对应的 AIC 的值，选择使得 AIC 最小的 λ 进行模型拟合。



使得 AIC 最小的 $\lambda = 0.0517$，进行模型拟合，从 200 个基因中选出了 4 个

变量，分别为：

gene_names	beta
Has267	8.2821856
Has1058	0.4559008
Has617	0.6040329
Has559	3.6396772

从表中可以看到，回归系数都是正的，表明这四个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。

6 lasso、MCP、L_0 penalt、group lasso 和 group MCP 结果总结

- 以上 5 种方法选出的基因没有重合；
- lasso、MCP、group lasso 和 group MCP 选出的基因也没有重合。
- lasso 和 MCP 选出了一个共同的基因，结果如下：

gene_names	lasso	MCP
Has1772	0.5847296	1.2002765
Has377	-0.6703120	-0.5840983
Has1870	0.4775175	0.5179080
Has617	-0.2614381	-0.2597950

回归系数为负值，表明肿瘤组织中该基因表达较少。

- group lasso 和 group MCP 方法均选出了 Has267、Has1058、Has559 三个基因，结果如下:

gene_names	group_lasso	group_MCP
Has267	0.6092363	8.2821856
Has1058	0.0045176	0.4559008
Has559	0.1724631	3.6396772

基因系数差异较大，但系数都为正，表明这三个基因在肿瘤组织表达更多，即其表达可能会促进肿瘤的生长。

- lasso、group lasso 和 group MCP 选出的基因中有一个相同的。

gene_names	lasso	group_lasso	group_MCP

可以看到回归系数不同，但 lasso、group lasso 的系数都很小，接近 0，group MCP 回归系数稍大，专业结论慎下。