LASSO 回归

LASSO，全称 Least Absolute Shrinkage and Selection Operator，是一种线性回归的缩减(shrinkage)方法。在如下两种情况时用得到：

1. 当自变量 X 本身存在线性相关关系（多重共线性）。如 X1 与 X2 非常相关（即使并不是完全线性相关），这时 X1 的回归系数 b1 与 X2 的回归系数 b2 就可以有无数种组合而得到完全相同的 b1*X1+b2*X2，也就是说对回归系数的求解不稳定，就需要附加条件来求解回归方程。
2. 当数据特征（X 变量数）比数据量（观察记录数）还要多的时候，也需要附加条件来求解回归方程。

LASSO 将回归系数（b）收缩在一定的区域内。LASSO 的主要思想是构造一个一阶惩罚函数获得一个精炼的模型，通过最终确定一些变量的系数为 0 进行特征筛选。LASSO 的惩罚项为：sum(abs(b)) <= t。

此模块调用 R glmnet 程序包进行 LASSO 回归分析，应变量类型可以是：0/1 两分类（binomial），正态连续性（gaussian），随时间依赖的 0/1 生存状态（survival），泊松分布（poisson）。自变量可以有各种类型（分类型，连续型），如果自变量是分类性，模块自动生成 0/1 哑变量带入模型筛选；如果自变量有缺失，自动生成一个 0/1 是否缺失的哑变量带入模型。

LASSO 常用在影像学数据的特征筛选，影像学测量特征之间存在很强的相关性，LASSO 方法可以从这些特征中筛选出一些有助于预测与诊断的特征。

此模块通过 10 folds 交叉验证（cross-validation）筛选 lambda。lambda 越大，模型越精简。交叉验证方法是将数据分成 10 等分，首先对全数据进行拟合，生成 lambda 序列，然后每次排除 1 分数据，用余下的 9 分数据进行验证，计算 10 次验证得出来的错误(deviance)的平均值与标准差。最终输出两种模型，一是基于 lambda.min 即错误的均值为最小时对应的 lambda；二是基于 lambda.1se 即错误均值在最小值的 1 个标准差范围之内对应的最大 lambda。

输出

(1) LASSO lambda 筛选图，上面标出 lambda.min 与 lambda.1se （Tuning parameter (lambda: λ) selection in the LASSO model used 10-fold cross-validation via minimum criteria）。输出作图所用数据 _cvm.xls 文件，以便用户重新制图。

(2) LASSO 回归系数与 lambda 对应关系图 （LASSO coefficient profiles of the features against the log(λ)）。输出作图所用数据 _coef.xls 文件，以便用户重新制图。

两个模型对应的回归系数。

两个模型对应的 ROC（当应变量为两分类与生存状态）与散点图（当应变量为连续性变量）。

下载练习数据：http://www.empowerstats.com/empowerStats/exdata/lassotest.xls

例 1：两分类应变量 LASSO 回归，应变量为 YBIN，自变量：X1-X30，输入界面如下：

输出结果：

Predict for: YBIN
Family: binomial
Predictors: X1; X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30

LASSO (Least Absolute Shrinkage and Selection Operator)
Tuning parameter(lambda) selection in the LASSO model used 10-fold cross-validation

Lambda.min (log) [value of lambda that gives minimum mean cross-validated error] : 0.0214 (-3.844)
Lambda.1se (log) [largest value of lambda such that error is within 1 standard error of the minimum] : 0.0411 (-3.1928)

Select lambda = lambda.1se: 0.0411 (-3.1928)
Variables selected: X2, X3, X4, X5, X6, X8, X9, X10, X22, X23, X25, X26, X28, X29
Formula for calculate score (not include Intercept): 0.3272*X2 - 0.22507*X3 - 0.74881*X4 - 0.07415*X5 - 0.50568*X6 - 0.26698*X8 + 0.31294*X9 - 0.83833*X10 + 0.12982*X22 + 0.13283*X23 +

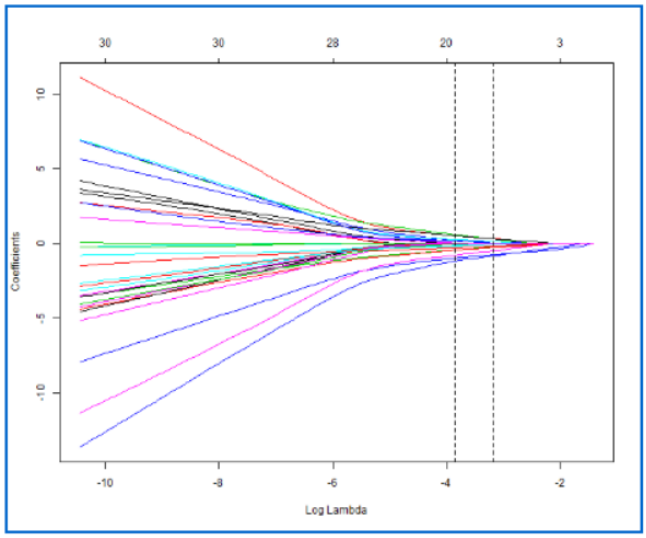
0.31878*X25 - 0.19038*X26 + 0.03177*X28 - 0.05855*X29

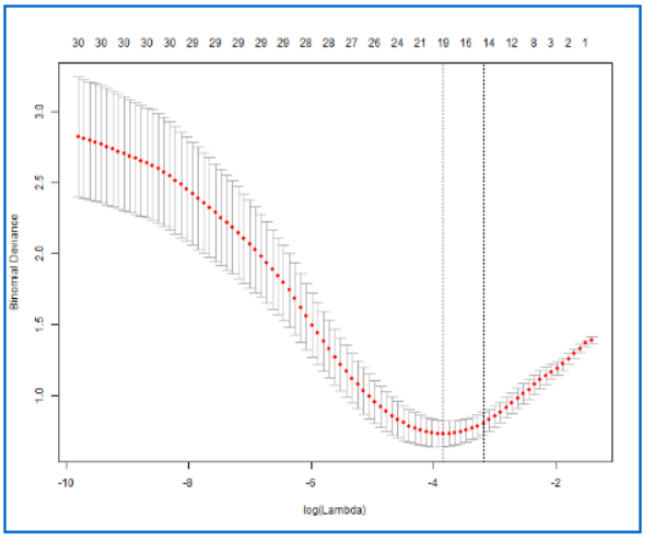
Select lambda = lambda.min: 0.0214 (-3.844)
Variables selected: X2, X3, X4, X5, X6, X8, X9, X10, X11, X12, X16, X22, X23, X25, X26, X27, X28, X29, X30
Formula for calculate score (not include Intercept): 0.52699*X2 - 0.43796*X3 - 0.98637*X4 - 0.13713*X5 - 0.78219*X6 - 0.44163*X8 + 0.60519*X9 - 1.19558*X10 - 0.02544*X11 - 0.01992*X12 + 0.22011*X16 + 0.17504*X22 + 0.27*X23 + 0.53944*X25 - 0.2876*X26 - 0.04983*X27 + 0.17122*X28 - 0.1731*X29 + 0.02248*X30

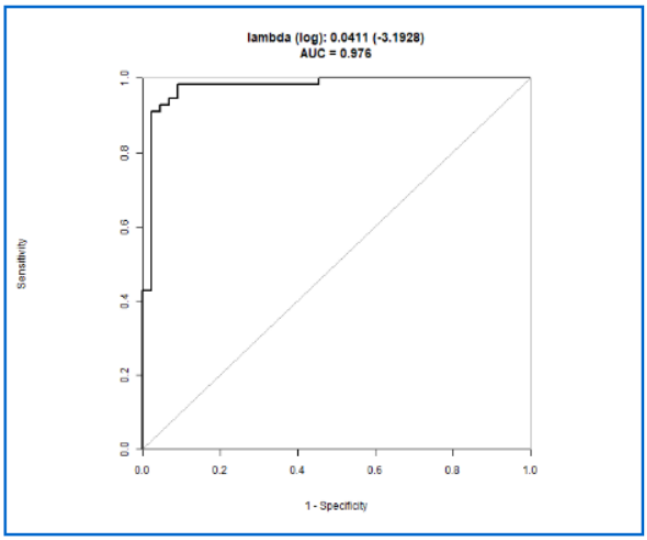
Coefficients

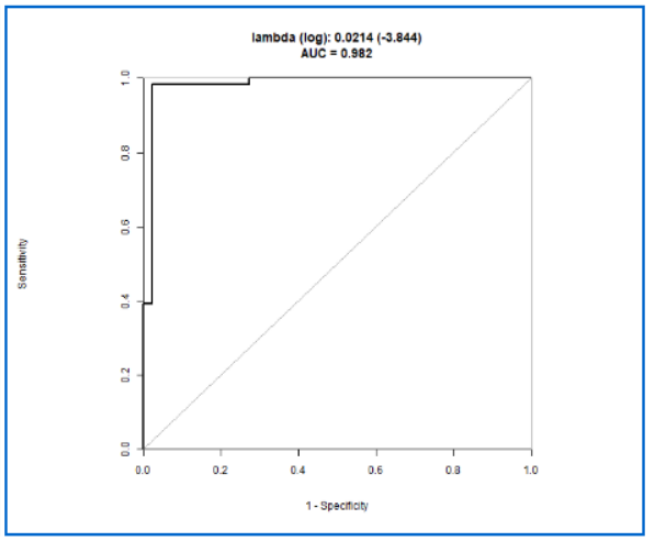
Lambda (log)	0.0411 (-3.1928)	0.0214 (-3.844)
(Intercept)	0.211622170092256	0.238908226332628
X1	0	0
X2	0.32720091639827	0.526990067314216
X3	-0.225067167533667	-0.4379629455825
X4	-0.748807576305804	-0.986367352059908
X5	-0.0741494050224252	-0.13713105225003
X6	-0.505676198314773	-0.782185972325462
X7	0	0
X8	-0.266976105733026	-0.441625708796562
X9	0.312939949745972	0.605191498869473
X10	-0.838330415209342	-1.19557557868958
X11	0	-0.0254383775986178
X12	0	-0.0199166155934526
X13	0	0
X14	0	0
X15	0	0
X16	0	0.220112424362534
X17	0	0
X18	0	0
X19	0	0
X20	0	0
X21	0	0
X22	0.129824741196536	0.175043700471903
X23	0.13282577676012	0.269999075916631
X24	0	0
X25	0.318776509289498	0.539444782500429
X26	-0.190378237165772	-0.287601429576103
X27	0	-0.0498304732186938
X28	0.0317750374376999	0.171224262601036
X29	-0.0585511125335968	-0.173105104328298

X30	0	0.022480452455188

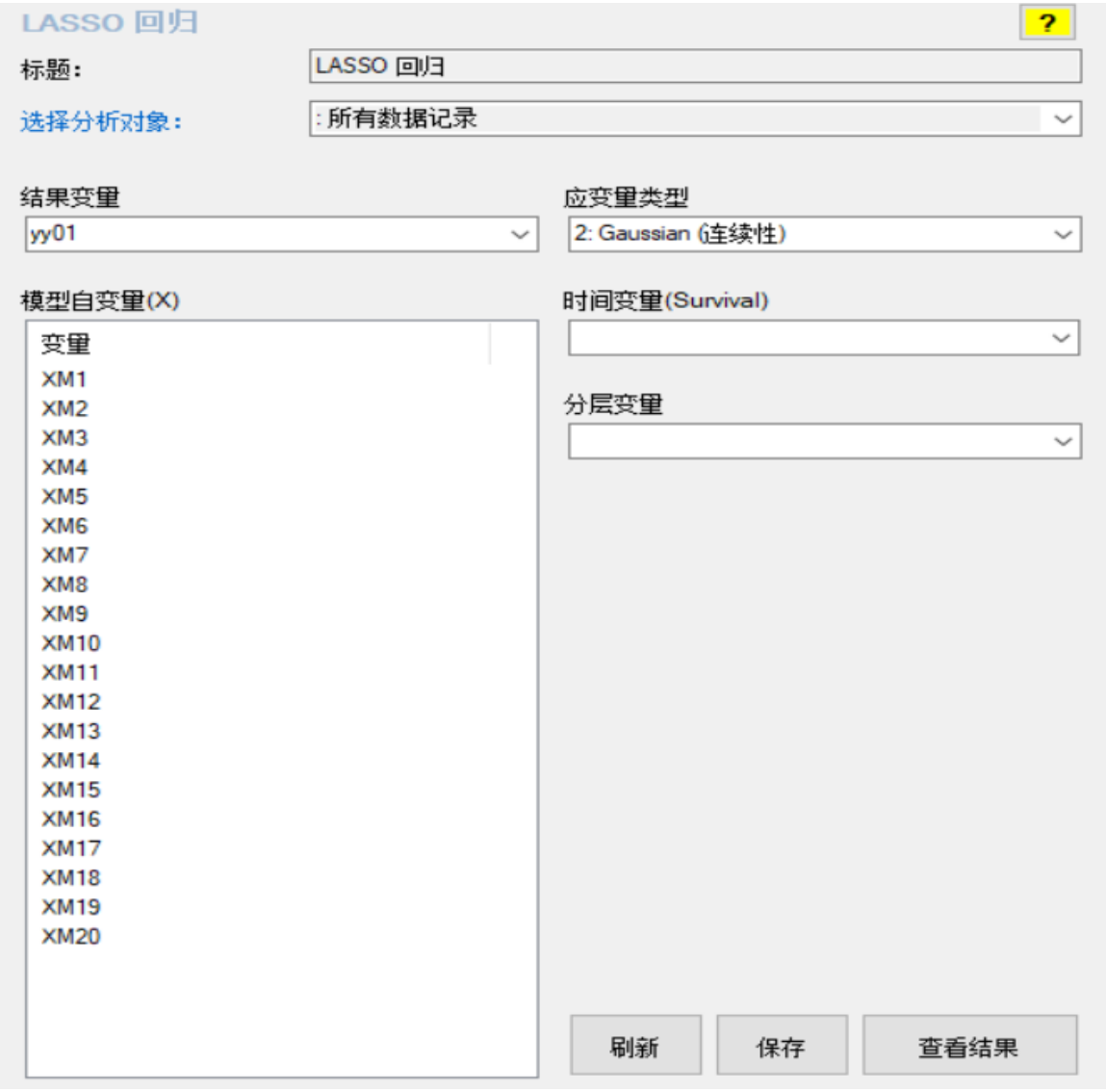








例 2：连续型应变量 LASSO 回归，应变量为 YY01，自变量：XM1-XM20，输入界面如下：



输出结果：

Predict for: YY01
Family: gaussian
Predictors: XM1; XM2; XM3; XM4; XM5; XM6; XM7; XM8; XM9; XM10; XM11; XM12; XM13; XM14; XM15; XM16; XM17; XM18; XM19; XM20

LASSO (Least Absolute Shrinkage and Selection Operator)
Tuning parameter(lambda) selection in the LASSO model used 10-fold cross-validation

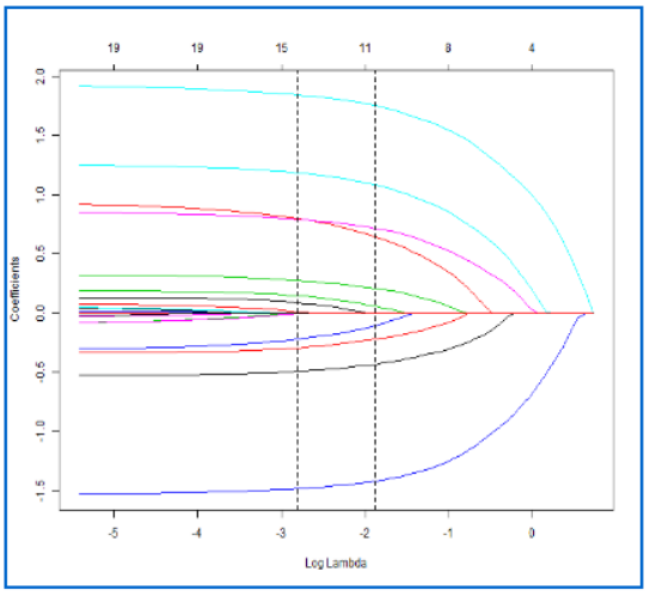
Lambda.min (log) [value of lambda that gives minimum mean cross-validated error] : 0.0602 (-2.81)
Lambda.1se (log) [largest value of lambda such that error is within 1 standard error of the minimum] : 0.1526 (-1.8797)

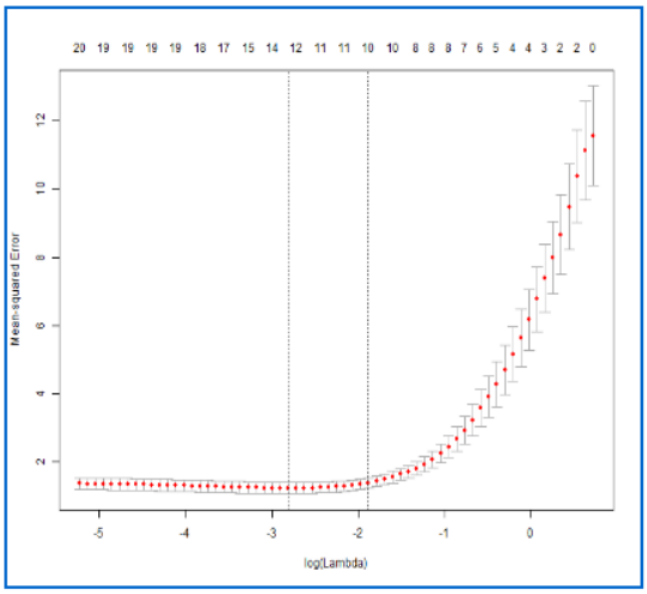
Select lambda = lambda.1se: 0.1526 (-1.8797)
Variables selected: XM1, XM2, XM5, XM9, XM10, XM11, XM12, XM14, XM15, XM16
Formula for calculate score (not include Intercept): - 0.43553*XM1 + 0.64517*XM2 + 1.08198*XM5 + 0.20564*XM9 - 1.42422*XM10 + 1.75311*XM11 + 0.71526*XM12 - 0.22442*XM14 + 0.05722*XM15 - 0.10767*XM16

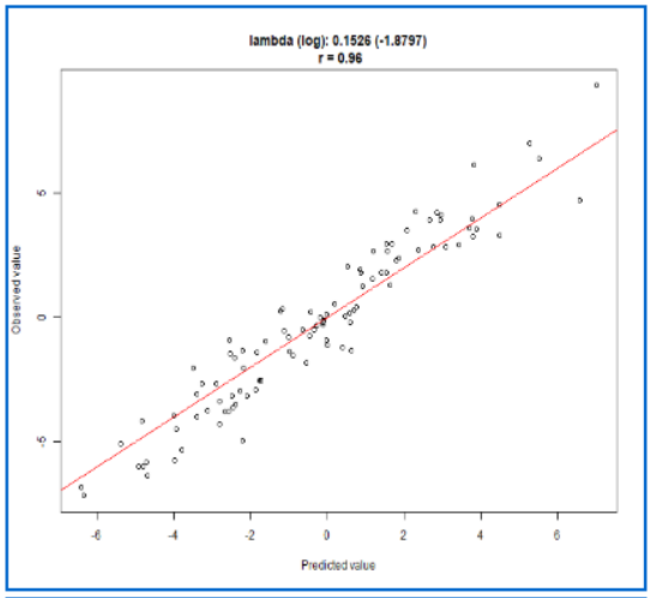
Select lambda = lambda.min: 0.0602 (-2.81)
Variables selected: XM1, XM2, XM5, XM9, XM10, XM11, XM12, XM13, XM14, XM15, XM16, XM18, XM20
Formula for calculate score (not include Intercept): - 0.49448*XM1 + 0.7973*XM2 + 1.1846*XM5 + 0.27213*XM9 - 1.48736*XM10 + 1.84567*XM11 + 0.7903*XM12 + 0.08431*XM13 - 0.29578*XM14 + 0.14432*XM15 - 0.22133*XM16 - 0.00012*XM18 + 0.00917*XM20

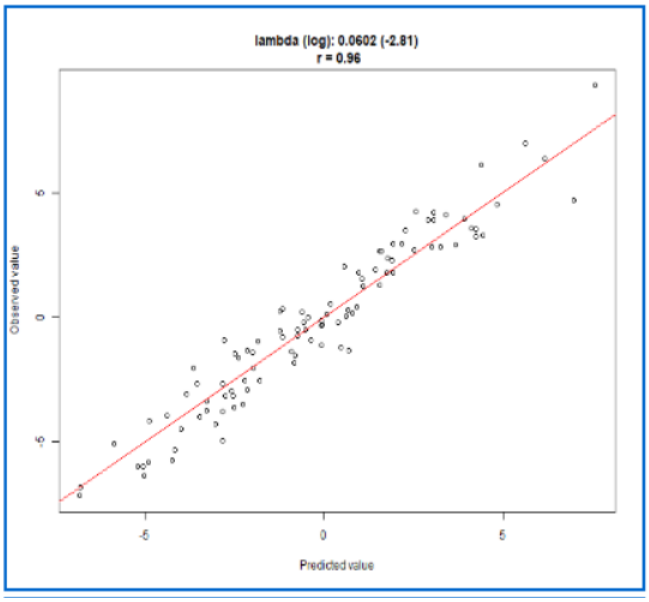
Coefficients

Lambda (log)	0.1526 (-1.8797)	0.0602 (-2.81)
(Intercept)	-0.177390371989018	-0.199409474896676
XM1	-0.435527596449108	-0.494482799464034
XM2	0.645173359197459	0.797296590248866
XM3	0	0
XM4	0	0
XM5	1.08198112590028	1.18459575097825
XM6	0	0
XM7	0	0
XM8	0	0
XM9	0.205644296510097	0.272132851709849
XM10	-1.42422434213449	-1.48736165215455
XM11	1.75310724862319	1.8456702167069
XM12	0.715259669038611	0.790296319838077
XM13	0	0.084308939830887
XM14	-0.224414521667208	-0.295782304431207
XM15	0.0572192751768561	0.144324444825295
XM16	-0.107668674037469	-0.221332551574302
XM17	0	0
XM18	0	-0.000116168252544583
XM19	0	0
XM20	0	0.00916909608881437


19 19 15 11 8 4
Coefficients
Log Lambda


20 19 19 19 19 18 17 15 14 12 11 11 10 10 8 8 8 7 6 5 4 4 3 2 2 0
Mean-squared Error
log(Lambda)


lambda (log): 0.1526 (-1.8797)
r = 0.96
Observed value
Predicted value


lambda (log): 0.0602 (-2.81)
r = 0.96
Observed value
Predicted value