Lasso 回归实例演示

张沥今

zhanglj37@mail2.sysu.edu.cn

为了验证传统 OLS 估计法容易出现过拟合的问题，展示 Lasso 回归的步骤和报告标准，促进 Lasso 回归的应用，本文将采用实例演示详细展示 Lasso 回归的分析流程，并对比传统估计方法。同时，实例分析还将纳入 Relaxed Lasso 方法。Relaxed Lasso 方法是指在采用 Lasso 回归选择合适的预测变量后，再采用 OLS 估计法针对这些预测变量重新建立回归模型。这种方法避免了传统 Lasso 方法压缩重要非零系数而容易产生估计偏差的问题(Serang et al., 2017)。分析采用 R 软件。

数据来源于 395 名葡萄牙中学生，数据中包含了 11 个连续变量：(1) 年龄(age)，(2) 家庭关系质量(famrel)，(3) 放学后空闲时间(freetime)，(4) 和朋友出去玩的频率(goout)，(5) 工作日饮酒频率(dalc)，(6) 周末饮酒频率(walc)，(7) 自评健康状况(health)，(8) 缺课次数(absences)，(9) 学生第一次数学测验成绩(G1)，(10) 中期测验成绩(G2)和 (11) 期末测验成绩(G3)。其中期末测验成绩为因变量，本研究将探究能够有效预测数学期末测验成绩的因素。相关分析结果显示，学生第一次数学测验成绩、中期测验成绩与期末测验成绩之间存在较强的正相关。

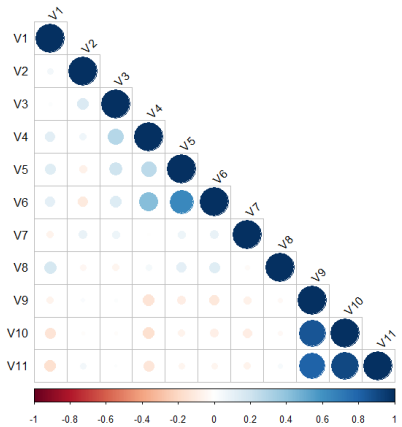

图 1 变量间相关图

注：红色系代表负相关，蓝色系代表正相关，颜色越深代表相关值越大。

在 Lasso 回归中，首先采用 10 重交叉验证方法选择合适的惩罚项 λ。这一方法可以通过 R 软件中的 glmnet 包(Friedman, Hastie, & Tibshirani, 2010)实现。值得注意的是，为了保

证每次交叉验证分析得到的λ结果一致，需要采用 set.seed()函数设定随机数种子，否则每次分析的结果会存在微小差异。

结果显示最小化均方误差(Mean Square Error, MSE)的λ为 0.043，λ+1se 为 0.776。图 2 呈现了随着 log(lambda)的增加 MSE 值的变化。当 λ 对复杂模型的惩罚力度增大时，MSE 同样会增大，而惩罚项的增大最终会导致所有系数压缩到 0，此时 MSE 值最大。

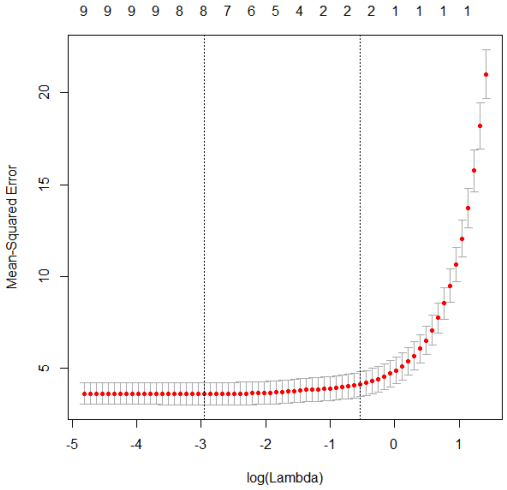

图 2 十重交叉验证结果

注：图中两条竖线分别代表最小化 MSE 的λ值和λ+1se 值

图 3 呈现了随着 log(lambda) 的增加，标准化回归系数被压缩的情况，可以看到的是，随着惩罚力度的增大，标准化系数最终全部会被压缩到 0。而在λ值为 0.776 处，有两个系数不为 0。根据输出结果，G1(学生第一次数学测验成绩)和 G2(学生中期数学测验成绩)两个预测因素被保留下来。

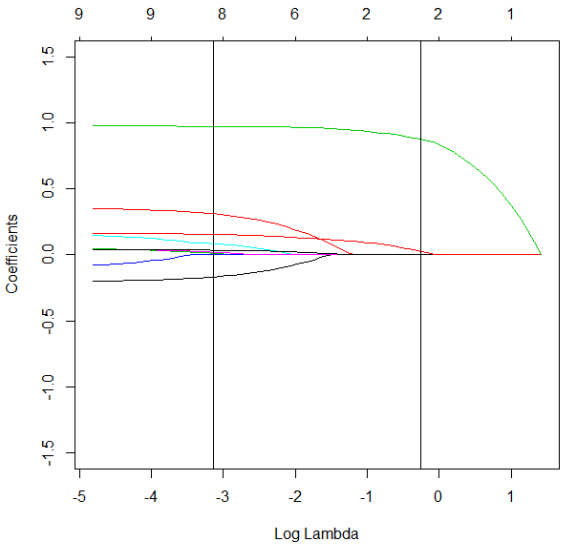

图 3 惩罚项对系数的压缩结果

此外，Lasso 回归中可以通过 covTest 软件包(Lockhart et al., 2014)计算参数估计的 p 值，进一步计算 p 值发现，同样只有 G1 和 G2 变量通过了显著性检验(表 1)。

而在 OLS 估计中，共发现了年龄、家庭关系质量、缺课次数，第一次测验成绩和期中成绩五个变量可以显著预测期末数学成绩(表 1)。但是结果显示缺课次数正向预测期末数学成绩，即学生缺勤次数越多，期末成绩越高($b = 0.042, p = 0.001$)，这显然和常识相悖。而相关分析也显示缺课次数和期末成绩间未发现显著相关(r = 0.034, p = 0.497)。而 OLS 回归分析得到的显著结果可能是由于样本量和观察指标数的比率较低(n / p = 3.95)，模型发生了过拟合现象，即模型在最小化结果变量的预测值和观测值的差异时，错误地学习到了不存在的规律。此外，和 Lasso 回归相比，OLS 额外发现的另外两个显著的预测变量和期末成绩的相关值较弱(图 1)。其中年龄和期末数学成绩显著负相关(r = -0.162, p = 0.001)，而家庭关系质量和期末数学成绩未发现显著相关($r = 0.051, p = 0.309$)。

进一步进行 Relaxed Lasso 分析，即采用 Lasso 回归选择出的 G1 和 G2 变量与期末数学成绩建立 OLS 回归模型。结果发现与传统的 OLS 估计相比，Relaxed Lasso 回归的 R 方、校正后 R 方及均方误差)均相差不大。即 Relaxed Lasso 回归仅采用两个预测变量就基本达到了 OLS 回归采用 5 个变量所获得的预测能力。

表 1 Lasso、 OLS、Relaxed Lasso 回归结果

	系数估计值(p 值)		
预测变量	OLS	Lasso	Relaxed Lasso

age	-0.206(0.009)**	-(0.072)	-
famrel	0.36(0.001)**	-(0.699)	-
freetime	0.058(0.57)	-(0.913)	-
gout	-0.014(0.891)	-(0.981)	-
dalc	-0.108(0.448)	-(0.646)	-
walc	0.17(0.105)	-(0.294)	-
health	0.046(0.509)	-(0.899)	-
absences	0.042(0.001)**	-(0.089)	-
G1	0.164(0.003)**	0.057(0.005)**	0.153(0.007)**
G2	0.977(<0.001)***	0.903(<0.001)***	0.987(<0.001)***
R^2	0.835	-	0.822
adjusted R^2	0.831	-	0.821
Mean Square Error	3.446	-	3.723

注：**代表 p 小于 0.01，***代表 p 小于 0.001。

从上述分析中可以看出，OLS 回归所选择的预测变量可能是不可靠且冗余的。一方面在本研究中 OLS 回归所选择的预测变量和因变量间相关很弱，另一方面，增加的三个预测变量并不能很好地提升对因变量的解释力，R 方和校正后 R 方的值都和仅采用两个预测变量的回归模型相接近。此外，Relaxed Lasso 方法也避免了 Lasso 方法在压缩不重要的系数的同时对非零系数(G1, G2)的压缩。

值得注意的是，Lasso 回归并不总是会获得更简洁的预测变量集，它的目的是采用较少的预测变量获得较高的预测能力。这尤其体现在样本量较少时，OLS 回归所使用的假设检验为了控制一类错误率，通常会获得较高的标准误估计，检验力较低，而 Lasso 回归在此时则更易于获得更高的检验力和预测能力。

Reference

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1-22.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*, 413–468.

Serang, S. , Jacobucci, R. , Brimhall, K. C. , & Grimm, K. J. . (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, *24.* 733-744.