

应用简报

基于Adaptive-Lasso的本科成绩统计分析 *

谭常春 张雪莲 胡俊迎

(合肥工业大学数学学院, 合肥, 230009)

摘要: 本文主要研究了合肥工业大学数学学院数学专业学生的本科成绩状况, 应用多元线性统计方法, 探讨了前期所有课程成绩对后期成绩的影响. 首先对前期课程进行主成分分析, 并采用逐步回归方法建立了后期成绩与主成分之间的回归模型. 其次, 采用Adaptive-Lasso方法建立后期成绩与前期课程成绩间的Adaptive-Lasso回归模型. 最后, 对以上模型进行对比分析. 研究表明, 基于Adaptive-Lasso方法的主成分回归模型能很好地拟合后期成绩, 并对后期成绩情况给出合理的解释.

关键词: 本科成绩; 主成分分析; 多元线性回归模型; Adaptive-Lasso

中图分类号: O212.4

§1. 引言

自全国高校扩招大学生以来, 如今社会中本科毕业生比比皆是, 因此对大学生的毕业质量就有了更高的要求. 目前, 教学已经成为了各大高校力抓的重点, 也是教育部门特别关注的方面. 本科在校成绩作为衡量大学生学习能力和解决问题能力的标准之一, 分析本科在校成绩并由此判断是否成为一名合格的毕业生变得尤为重要.

有关学业状况的统计分析, Frisbee^[1]研究学生资质、老师与课程特点以及学生的时间分配等因素对大学课程成绩的影响, 建立线性模型, 并采用两阶段最小二乘法进行估计. Rosander^[2]基于纵向研究的设计方法, 探讨了学生的个性、智商以及学习方法的重要性, 阐述了大学生的学习状况不是受单方面因素影响, 而是受多元因素影响. Onyper等人^[3]采用路径分析法, 对大学生的课程开始时间、睡眠和学习成绩进行了相关研究, 分析了大学生的在校学习及生活状况. 面对中国体制中诸多的教育教学问题, 国内不少学者也对高校学生的学业状况展开了一系列研究. 丁澍和缪柏其^[4]对本科生大学期间各个学期和各种类型课程的成绩分别进行了因子分析、聚类分析及multinomial logistic回归分析, 研究了大学成绩的特点及其影响因素. 鲁威等人^[5]根据上海交通大学医学系的高考成绩及其本科医学课程成绩, 采用回归统计对其进行了分析, 这项研究在教育教学和学生管理方面都起到了辅助作用. 孙毅等人^[6]基于多元线性回归模型, 解析了国家四级英语考试成绩与学生期末考试成绩间的关系, 对考试成绩做出了合理评价与预测.

*全国统计科研计划重点项目(2012LZ009)和中央高校基本科研业务费(JZ2015HGJX0177, JS2016HGJ0019)资助.
本文2015年9月30日收到.

在学业状况研究中, 通常采取建立线性模型的方法, 并用普通最小二乘法求解模型的参数估计. 但对很多情况, 比如变量间存在严重的多重共线性, 这种方法的估计值并不准确, 从而导致预测精度较差. 此外, 自变量数量较多时, 模型的解释也就变得困难. 改进的最小二乘估计方法有子集选择和岭回归法, 前者的模型解释性好但模型变得不稳定, 后者与之相反. Tibshirani^[7]提出了Lasso方法, 这种方法的目的是为了压缩模型参数, 使得某些回归系数变小甚至为0, 兼具子集选择和岭回归的优点. 而后, 无数学者提出了Lasso方法的改进技术. Zou^[8]指出Lasso估计对所有系数的压缩程度相同, 因此他提出了Adaptive-Lasso方法, 这种方法对不同系数使用不同程度的压缩效果, 使Lasso估计具有Oracle性质. 如今在许多实际问题中, 应用Adaptive-Lasso方法进行多元统计分析取得了很好的成绩. 谭常春等人^[9]针对城市火灾次数与气象因素的Adaptive-Lasso分析, 很好地预测了城市火灾次数.

本文针对合肥工业大学数学学院本科生在校成绩, 分析本科课程对本科成绩的影响, 在主成分分析的基础上运用Adaptive-Lasso方法进行建模, 并预测本科后期成绩. 不仅有助于我们了解当代大学生学习特点, 同时也能帮助教学工作者更好地安排教学内容, 以及为同学们的学习方向提供一定参考.

§2. 预备知识

2.1 主成分模型

主成分分析就是利用降维思想, 在数据信息损失量最少的原则下, 将多变量进行最佳综合和简化的方法, 其模型为

$$\left\{ \begin{array}{l} Y_1 = a_1^\top X = a_{11}X_1 + a_{21}X_2 + \cdots + a_{m1}X_m \\ Y_2 = a_2^\top X = a_{12}X_1 + a_{22}X_2 + \cdots + a_{m2}X_m \\ \vdots \\ Y_m = a_m^\top X = a_{1m}X_1 + a_{2m}X_2 + \cdots + a_{mm}X_m \end{array} \right.,$$

其中 X_1, X_2, \dots, X_m 为感兴趣的 m 个解释变量, Y_1, Y_2, \dots, Y_m 为生成的 m 个主成分. 记 $Y = (Y_1, Y_2, \dots, Y_m)^\top$, $X = (X_1, X_2, \dots, X_m)^\top$, $a_i = (a_{1i}, a_{2i}, \dots, a_{mi})^\top$, 并满足 $a_i^\top a_i = 1$, $i = 1, 2, \dots, m$. 实际问题中, 主成分 Y 的维度比 X 小, 即 $Y = (Y_1, Y_2, \dots, Y_p)^\top$ ($p < m$), p 的选择须使得主成分贡献率不低于 70%.

所得 $Y_1 = a_1^\top X$ 被称为 X_1, X_2, \dots, X_m 的第一主成分, 要求 Y_1 尽可能反映原 m 个变量的信息. 这里, “信息”用方差来度量, 即 $\text{Var}(Y_1)$ 越大, 表示 Y_1 含 X_1, X_2, \dots, X_m 中的信息就越多. 若 Y_1 不足以反映原变量的“信息”, 再考虑求出第二主成分 $Y_2 = a_2^\top X$, 并满足 $\text{Cov}(Y_1, Y_2) = a_1^\top \Sigma a_2 = 0$ (Σ 为 X 的协方差矩阵). 所得的 m 个主成分 Y_1, Y_2, \dots, Y_m 彼此间互不相关, 这样避开了变量间共线性的问题.

主成分求法: 先求出 X 的协方差阵 Σ 的特征值-特征向量对 (λ_i, e_i) , $i = 1, 2, \dots, m$, 且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$, 则第*i*个主成分由 $Y_i = e_i^\top X$ 给出.

2.2 Adaptive-Lasso方法

Lasso是一种参数估计和变量选择同步的技术, 其参数估计定义为

$$\hat{\beta}_{\text{Lasso}} = \arg \min \left\| Y - \sum_{j=1}^p X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p |\beta_j|. \quad (1)$$

式(1)中, λ 是非负正则化参数, $\beta = (\beta_1, \beta_2, \dots, \beta_p)^\top$ 是回归系数, $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^\top$, $j = 1, 2, \dots, p$ 是预测变量, $X = (X_1, X_2, \dots, X_p)$ 是预测变量矩阵, $Y = (Y_1, Y_2, \dots, Y_n)^\top$ 是响应变量. 式1的第二部分称为“ l_1 惩罚”, 随着 λ 的增加, Lasso方法使得系数连续地趋于0, 若 λ 足够大时, 则系数缩小到0.

一个改进的Lasso方法, 称为Adaptive-Lasso方法, 其参数估计定义为

$$\hat{\beta}^{*(n)} = \arg \min \left\| Y - \sum_{j=1}^p X_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j|. \quad (2)$$

式(2)中, $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$ ($\gamma > 0$), $j = 1, 2, \dots, p$, 其中 $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)^\top$ 为普通最小二乘法所得系数估计值. 记权重向量

$$\hat{W} = (\hat{w}_1, \hat{w}_2, \dots, \hat{w}_p)^\top = \left(\frac{1}{|\hat{\beta}_1|^\gamma}, \frac{1}{|\hat{\beta}_2|^\gamma}, \dots, \frac{1}{|\hat{\beta}_p|^\gamma} \right) = \frac{1}{|\hat{\beta}|^\gamma} \quad (\gamma > 0).$$

§3. 主成分分析

本文中所使用的成绩数据来自合肥工业大学教务部, 为06至09级数学与应用数学(简称“数学”)专业, 共4个年级的学生在本科前六个学期中的各课程考试成绩(所用数据经过标准化处理). 首先对06–09级数学专业前5个学期所有课程的进行主成分分析, 并建立回归模型.

3.1 主成分分析

以06级为例, 先对全体学生前5个学期的课程成绩进行是否适合主成分分析的检验, 采用KMO检验和Bartlett球体检验, 结果如表1所示:

表1 KMO检验和Bartlett球体检验

KMO检验		0.866
Bartlett球体检验	近似卡方	2 458.941
	自由度	780
	显著值	0.000

表1显示KMO检验的值为0.866, 且Bartlett球体检验的P值为0.000, 因此合适进行主成分分析. 采取最大方差因子旋转对成分矩阵进行旋转, 最终的主成分分析结果如表2及表3所示.

表2 解释的总方差

成分	初始特征值			提取平方和		
	合计	方差的%	累积%	合计	方差的%	累积%
1	15.366	38.416	38.416	15.366	38.416	38.416
2	3.169	7.922	46.337	3.169	7.922	46.337
3	2.094	5.235	51.572	2.094	5.235	51.572
4	1.770	4.426	55.998	1.770	4.426	55.998
5	1.478	3.696	59.694	1.478	3.696	59.694
6	1.353	3.383	63.077	1.353	3.383	63.077
7	1.341	3.352	66.429	1.341	3.352	66.429
8	1.176	2.940	69.369	1.176	2.940	69.369
9	1.049	2.622	71.991	1.049	2.622	71.991

表3 旋转成份矩阵

	成份								
	1	2	3	4	5	6	7	8	9
数分A	0.786	0.296	0.056	0.038	0.156	0.089	-0.071	0.045	-0.037
高代A	0.869	0.096	-0.063	0.025	0.061	-0.029	0.023	0.122	0.105
解几	0.860	-0.013	0.157	-0.132	0.062	-0.059	0.148	-0.055	-0.046
计算机	0.282	0.199	0.016	0.151	0.727	0.045	0.248	0.014	0.083
思修	0.254	0.127	0.549	0.135	0.223	0.226	-0.012	0.256	-0.125
体育1	-0.239	0.093	0.294	-0.165	0.280	0.367	0.535	0.070	-0.016
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

由表2可知, 从40门课程(原变量 X)中提取了9个主成分, 其累计贡献率达到71.991% (本文提取主成分的标准为: 贡献率> 70%).

由表3可知, 第一主成分主要为校定专业必修课(数学分析、解析几何、常微分方程等)及院定数学基础课(高等代数、复变函数、概率论等), 其贡献率为38.416%; 第二主成分主要为院定专业选修课(微分几何、数据库原理、西方经济学等), 其贡献率为7.922%; 第三至第五个主成分主要为校定公共课(形势与政策、军事理论、中国近代史等), 其累计贡献率为13.357%; 第六至第八个主成分主要为校定体育及实践课(体育和军事训练), 其贡献率累计为9.675%; 第九个主成分主要为校定英语课(大学英语课程), 其贡献率为2.622%.

按照同样的方法, 依次可得07-09级数学专业的主成分分析结果.

07数学: 第一主成分主要为院定数学基础课(实变函数、复变函数、概率论等)及院定专业选修课(数据结构、常用数学软件、西方经济学等), 其贡献率为33.623%; 第二个主成分主要为校定专业必修课(数学分析、解析几何、面向对象程序设计等), 其贡献率为7.086%; 第三个主成分主要是校定英语课(大学英语), 其贡献率为5.687%; 第四至第六及第九个主成分主要为校定公共课(形势与政策、军事理论、中国近代史等), 其贡献率累计为14.887%; 第七至第八及第十个主成分主要为校定体育及实践课(体育和军事训练), 其贡献率累计为8.747%.

08数学: 第一主成分主要为院定数学基础和必修课(常微分方程、复变函数、概率论等), 其贡献率为29.656%; 第二和第六个主成分主要为校定公共课(形势与政策、军事理论、思想道德等), 其贡献率为16.762%; 第三至第四及第七个主成分主要为校定数学专业必修课(数学分析、解析几何、大学物理等)及院定专业选修课(离散数学、微分几何、西方经济学等), 其累计贡献率为14.759%; 第五主成分主要为校定英语课(大学英语), 其贡献率为4.465%; 第八及第九个主成分主要为校定体育及实践课(体育和军事训练), 其贡献率累计为5.795%.

09数学: 第一主成分主要为院定数学基础和必修课(高等代数、常微分方程、概率论等)及校定英语课(大学英语), 其贡献率为34.677%; 第二主成分主要为校定数学专业必修课(解析几何、面向对象程序设计等), 其贡献率为7.692%; 第三至第五、第七至第八及第十个主成分主要为校定公共课(形势与政策、军事理论、思想道德等), 其累计贡献率为23.679%; 第六和第九个主成分主要为校定体育及实践课(体育和军事训练), 其贡献率累计为6.483%.

由于06级和07级执行05版教学计划, 而08级和09级执行08版教学计划, 因此, 各年级主成分的选定在个数和含义上略有差异. 但两版教学计划在课程的类型和课程的时间设置上变动较小, 因此各年级的主成分个数及含义基本相似.

3.2 建立回归模型

利用主成分方法所得的成分矩阵, 采用逐步回归方法建立第六学期总体成绩(y)与前五学期主成分(prin1 ~ prin9)之间的回归模型(以06级为例), 所得结果如表4:

表4 回归系数

	估计值	标准误差	T值	显著值
截距项	78.719	0.447	176.212	0.000
prin1	1.483	0.115	12.935	0.000
prin4	-0.992	0.338	-2.937	0.004
prin5	-1.448	0.369	-3.918	0.000

方程F检验的P值: 0.000.

表4给出了模型的回归系数及相应的检验。模型通过了F检验，各回归系数也通过了T检验，拟合优度为0.697，均方误差为17.362，所得预测模型为

$$y = 78.719 + 1.483 \times \text{prin1} - 0.992 \times \text{prin4} - 1.448 \times \text{prin5}.$$

类似可得到07-09级数学专业的预测模型，总结如表5所示：

表5 06-09主成分逐步回归结果

年级	回归方程	R^2	MSE
06数学	$y = 78.719 + 1.483 \times \text{prin1} - 0.992 \times \text{prin4} - 1.448 \times \text{prin5}$	0.697	17.362
07数学	$y = 69.694 + 2.798 \times \text{prin1}$	0.583	83.887
08数学	$y = 77.966 + 2.544 \times \text{prin1} + 1.374 \times \text{prin3}$	0.725	30.201
09数学	$y = 77.998 + 1.737 \times \text{prin1} - 1.527 \times \text{prin2}$	0.518	46.925

由表5知，模型的拟合优度并不高，因此，考虑基于Adaptive-Lasso方法建立回归模型。

§4. Adaptive-Lasso

4.1 Adaptive-Lasso回归模型

由于Adaptive-Lasso是一种参数估计和变量选择同步的技术，在本节拟建立第六学期总体成绩(y)与前5个学期所有课程成绩(X_1, X_2, \dots, X_{40})间的Adaptive-Lasso回归模型。所得模型如表6(以06级为例)：

表6 回归系数表

截距项	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
36.05248	0	0	0	0	0	0	0	0	0	0
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}
0	0	0	0	0	0	0	0.0963	0	0.1037	0
X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}	X_{29}	X_{30}	X_{31}	X_{32}
0	0	0	0	0	0	0	0	0	0	0
X_{33}	X_{34}	X_{35}	X_{36}	X_{37}	X_{38}	X_{39}	X_{40}			
0	0	0	0	0.3398	0	0	0			

由表6可知，Adaptive-Lasso方法保留了第18、20和37门课程，即数学分析(III)、常微分方程(II)和数据库原理。所得模型的拟合优度为0.83，均方误差为18.46349。因此预测模型为

$$y = 36.05248 + 0.0963 \times X_{18} + 0.1037 \times X_{20} + 0.3398 \times X_{37}.$$

类似可得到07–09级的Adaptive-Lasso预测模型, 如表7:

表7 06–09 Adaptive-Lasso回归结果

年级	回归方程	R^2	MSE
06数学	$y = 36.05248 + 0.0963 \times X_{18} + 0.1037 \times X_{20} + 0.3398 \times X_{37}$	0.83	18.46349
07数学	$y = 18.20753 + 0.2166 \times X_{13} + 0.2806 \times X_{20} + 0.2753 \times X_{41}$	0.828	68.97871
08数学	$y = -78.57086 + 0.0778 \times X_{15} + 0.2281 \times X_{19} + 0.1357 \times X_{22}$ $+ 0.6642 \times X_{24} + 0.5955 \times X_{28} + 0.048 \times X_{29}$ $+ 0.0964 \times X_{33} + 0.0548 \times X_{34}$	0.834	28.47147
09数学	$y = 48.02478 + 0.3941 \times X_{35}$	0.803	47.39269

由于初始变量较多且彼此共线性问题严重, Adaptive-Lasso方法选择的因变量不具足够的代表性, 为此在下一节中基于所选择的主成分进行Adaptive-Lasso建模.

4.2 基于Adaptive-Lasso的主成分回归模型

本节拟建立第六学期总体成绩(y)与前5个学期的主成分(prin1~prin9)间的Adaptive-Lasso回归模型, 所得结果如表8所示(以06级为例).

表8 回归模型

	估计值	标准误差	T值	P值	显著性
截距项	78.71897	0.4738641	166.121417	0.000000000	***
prin1	1.451765	0.1215778	11.941040	0.000000000	***
prin4	-0.5934945	0.3582777	-1.656521	0.101392241	
prin5	-1.1207356	0.3918903	-2.859820	0.005359963	**

方程F检验的P值: 0.000.

由表8可以看到Adaptive-Lasso回归模型选取了prin1、prin4、prin5这三个成分, 主要包括数学基础课、数学必修课和政治素养课. 而prin2、prin3及prin6~prin9主要包括选修课和校内通识课, 它们与第六学期成绩间的相关性很小以致可以忽略. 主成分方法解决了变量间的共线性, 而Adaptive-Lasso在此基础上剔除了对第六学期成绩影响微小的因素.

另外, 模型整体是非常显著的, 拟合优度为0.731, 均方误差为17.83199, 因此预测模型为

$$y = 78.71897 + 1.4518 \times \text{prin1} - 0.5935 \times \text{prin4} - 1.1207 \times \text{prin5}.$$

类似可得到07–09级的主成分Adaptive-Lasso回归模型, 如表9:

表9 06-09主成分Adaptive-Lasso回归结果

年级	回归方程	R^2	MSE
06数学	$y = 78.71897 + 1.4518 \times \text{prin1} - 0.5935 \times \text{prin4} - 1.1207 \times \text{prin5}$	0.731	17.83199
07数学	$y = 69.69385 + 2.737849 \times \text{prin1}$	0.605	83.94093
08数学	$y = 77.96578 + 2.495618 \times \text{prin1} + 1.02984 \times \text{prin3}$	0.745	30.56179
09数学	$y = 77.99848 + 1.638958 \times \text{prin1} - 1.024039 \times \text{prin2}$	0.587	47.86493

4.3 模型比较

本文主要运用主成分逐步回归、Adaptive-Lasso回归及主成分Adaptive-Lasso回归三种方法，对06-09数学专业第六学期总体成绩进行建模分析，各方法的 R^2 和MSE如表10：

表10 模型比较

年级	主成分逐步		Adaptive-Lasso		主成分Adaptive-Lasso	
	R^2	MSE	R^2	MSE	R^2	MSE
06数学	0.697	17.362	0.830	18.46349	0.731	17.83199
07数学	0.583	83.887	0.828	68.97871	0.605	83.94093
08数学	0.725	30.201	0.834	28.47147	0.745	30.56179
09数学	0.518	46.925	0.803	47.39269	0.587	47.86493

由表10可知，对原始变量直接建立Adaptive-Lasso回归模型，具有更高的拟合优度，但部分模型所选择的自变量较多且彼此间具有共线性问题，所以从实际角度出发，主成分Adaptive-Lasso回归所得模型相对稳定且模型解释性更好。

利用主成分Adaptive-Lasso回归方法，对06数学专业第六学期成绩进行拟合，如图1所示，拟合效果较优。

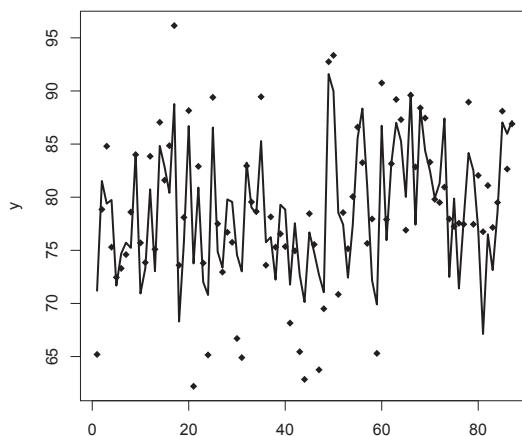


图1 模型预测

根据06–09级数学专业的主成分Adaptive-Lasso回归模型结果可知: 第一, 影响本科后期成绩最大的课程是院定数学基础及必修课, 其次是校定专业必修课; 第二, 06至07级的影响因素波动较大, 而08至09级的影响因素趋于平稳; 第三, 各年级后期成绩的预测精度明显不同. 主要原因有以下几点: 首先, 数学专业的学生主攻数学, 有明显的专业偏向; 其次, 06–07级和08–09级执行的教学计划版本不同, 且07级第六学期的课程设置与06级差别较大, 而08级与09级只是微差; 再者, 一些院定专业课程的师资变动, 在一定程度上导致了各年级的后期成绩波动较大, 进而各年级的预测效果不尽相同.

§5. 总 结

本文对合肥工业大学数学学院06–09级数学专业前六学期课程及其成绩, 分别建立了主成分逐步回归模型、Adaptive-Lasso回归模型以及主成分Adaptive-Lasso回归模型. 三者对比分析后可得, 主成分逐步回归虽保证了较好的模型解释和预测精度, 但模型的拟合优度较差; 而基于Adaptive-Lasso的主成分回归则在前者基础上改进了拟合优度, 提高了预测工作的准确性和实际性; 最后对原始变量直接进行Adaptive-Lasso回归, 虽满足了较高的拟合优度且预测精度也较好, 但从实际角度考虑, 其模型解释性不高, 没有很好的实际意义. 结果表明, 对于有专业偏向的学生, 其专业基础课及必修课对后期成绩有显著的影响. 由预测效果可知, 模型还存在一定的偏差, 主要原因是本科课程并不是影响本科成绩的唯一因素, 还包括学生自身素质及生活环境等诸多因素.

参 考 文 献

- [1] Frisbee W R. Course grades and academic performance by university students: a two-stage least squares analysis [J]. *Res. High. Educ.*, 1984, **20**(3): 345–365.
- [2] Rosander P. *The Importance of Personality, IQ and Learning Approaches: Predicting Academic Performance* [M]. Sweden: Lund University Publications, 2012.
- [3] Onyper S V, Thacher P V, Gilbert J W, et al. Class start times, sleep, and academic performance in college: a path analysis [J]. *Chronobiol. Int.*, 2012, **29**(3): 318–335.
- [4] 丁澍, 缪柏其. 当今本科生学业状况的统计分析 [J]. 中国科学技术大学学报, 2010, **40**(6): 557–564.
- [5] 鲁威, 杨云, 张剑戈, 等. 基于高考和医学课程成绩的医学生学业潜力的研究 [J]. 上海交通大学学报(医学版), 2012, **32**(10): 1373–1377.
- [6] 孙毅, 刘仁云, 王松, 等. 基于多元线性回归模型的考试成绩评价与预测 [J]. 吉林大学学报(信息科学版), 2013, **31**(4): 404–408.
- [7] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. *J. Roy. Statist. Soc. Ser. B*, 1996, **58**(1): 267–288.
- [8] Zou H. The adaptive Lasso and its oracle properties [J]. *J. Amer. Statist. Assoc.*, 2006, **101**(476): 1418–1429.

- [9] 谭常春, 谭景宝, 朱华亮. 城市火灾次数与气象因素的Adaptive-Lasso分析 [J]. 应用数学与计算数学学报, 2013, **27(3)**: 408–414.
- [10] 何晓群, 刘文卿. 应用回归分析 [M]. 3版. 北京: 中国人民大学出版社, 2011.
- [11] 梅长林, 周家良. 实用统计方法 [M]. 北京: 科学出版社, 2002.

The Statisitical Analysis of Undergraduate Grades Based on the Adaptive-Lasso

TAN Changchun ZHANG Xuelian HU Junying

(School of Mathematics, Hefei University of Technology, Hefei, 230009, China)

Abstract: In this paper, the multivariate linear statistical method is applied to research the undergraduate grades of students from the school of mathematics in Hefei University of Technology, and explore the impact on the later achievement by the early stage of achievement from all undergraduate courses. First, we get the main components from the previous courses by principal component analysis, then construct a linear regression model between the later achievement and main components by the stepwise regression method. Next, a linear regression model between the later achievement and the early stage of achievement from all undergraduate courses is constructed by Adaptive-Lasso method. Finally, comparative analysis is performed for the result of the above models. The research shows that the principal component regression model based on the Adaptive-Lasso method can well fit the later achievement, and give a reasonable explanation for the later academic performance.

Keywords: undergraduate grades; principal component analysis; multivariate linear regression model; Adaptive-Lasso

2010 Mathematics Subject Classification: 62H25; 62J05; 62P99