鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2022,38(4):350—356

基于多元线性回归和 Lasso 回归的高校生源质量影响因素研究

李学泱,邵喜高

(鲁东大学 数学与统计科学学院,山东 烟台 264039)

摘要: 为进一步提高高校生源质量,本文选取可能影响生源质量的多个因素,并从权威数据网站上获取所需数据用于研究分析。首先对高校生源质量与各因素作相关性分析,并应用 SPSS 软件建立多元线性回归模型;其次,对模型进行显著性检验,并运用逐步回归法进行变量选择,最终得到“最优”多元线性回归模型。基于上述数据,运用 Lasso 回归进行变量选择,检验经过筛选处理后变量的显著性,最终确定对高校生源质量有显著影响的因素。本文所用两种方法得到的结果均表明:本科生人数、重点学科、综合声誉指数以及高校是否属于理工类学校对高校生生源质量具有显著影响。

关键词: 多元线性回归;Lasso 回归;显著性检验;相关性

中图分类号: O213 **文献标志码:** A **文章编号:** 1673-8020(2022)04-0350-07

近年来,随着我国综合国力以及国家文化地位的不断提高,我国在高校生源质量的要求方面也日趋严格[1],这不仅仅关系着我国国民素质的整体水平,更关乎我国的国际地位。随着我国高等教育大众化逐步深入,高等学校之间在生源上的竞争亦是愈发激烈[2]。高校生源质量是高校进行人才培养的重要前提和基础,同时也对高校人才培养的起点和高度起着基础建设作用。2006 年教育部提出了“稳定规模,提高质量”的招生准则,同时要求高校招生的扩招规模要做到逐年缩减。2007 年,教育部副部长袁仁贵指出“针对当前的形势,应该进一步把握好规模发展节奏,切实把高校教育发展重点放在提高质量上”;史秋衡等[2]使用国家大学生学习情况调查的数据,分析不同类型高校生源质量的公平性和适切性;赵良君等[3]从高校的综合实力、知名度、就业前景、地域条件和招生宣传等八个方面分析了高校生源质量的影响因素,提出了吸引优秀生源的措施和对策,论述主要来源于新闻资料报道,缺乏实际数据的支撑;侯爽等[4]针对高校的声誉、专业实力、高校的办学地区和高校的专业前景等因素对高校生源质量的影响以及生源质量不同的高校就业现状和存在的问题进行了研究,但是研究的影响因素较少,研究重点在于生源质量与就业现状;赵飞等[5]对 985 高校本科生生源质量评价及影响因素进行分析,虽然主要基于高考分数来判定高校生源质量,但其中重点提到高校所处城市和地理位置等因素对生源质量的影响,这给本文影响因素的选择提供了思路。

通过调查研究影响高校生源质量的诸多因素,本文重点分析理工类、综合类、科学研究指数、一线城市、沿海城市、高校院士数、综合声誉指数、本科生人数、高校本科就业率、高校是否位于北京上海、社会服务指数、高校师生比、高校重点学科等因素。通过多元线性回归和 Lasso 回归建立合适的模型,分析各个因素对高校生源质量的影响程度,对其中的影响因素进行调控以达到对高校生源质量的有效控制,最终达到帮助高校找出可控因素、提高高校生源质量的目的[6]。

1 数据来源与获取及相关符号说明

本文所使用的数据来自 2021 年软科中国大学排名以及选取的 32 所 985、211 高校官方网站。

收稿日期: 2021-12-28;**修回日期:** 2022-06-20

基金项目: 山东省社会科学规划研究项目(20CSDJ10)

第一作者简介: 李学泱(1997—),女,山西阳泉人,硕士研究生,研究方向为应用统计、纵向数据分析。E-mail:ldlixy@163.com

通信作者简介: 邵喜高(1976—),男,山东莱阳人,硕士研究生导师,博士,研究方向为应用统计、纵向数据分析。E-mail:853395913@qq.com

其中,科学研究指数、高校院士数、综合声誉指数、高校重点学科这 4 个指标的数据来源于软科中国大学排名,本科生人数、高校本科就业率以及师生比所需数据主要从高校官方网站获取,重点参考了 2021 年高校相关数据。另外,科学研究指数是以学科基础、理科科研产出指数和文科科研产出指数为指标计算得出; 社会服务指数是理科社会服务指数和文科社会服务指数的加权平均; 综合声誉的计算以同行声誉、媒体声誉和国际声誉为基础。

为了便于下文的论述和写作,对高校生源质量指数和理工类、科学研究指数等影响因素作出规定: Y 代表高校生源质量指数,X_1 代表理工类,X_2 代表综合类,X_3 代表科学研究指数,X_4 代表一线城市,X_5 代表沿海城市,X_6 代表高校院士数,X_7 代表综合声誉指数,X_8 代表本科生人数,X_9 代表高校本科就业率,X_{10} 代表高校是否位于北京上海,X_{11} 代表社会服务指数,X_{12} 代表高校师生比,X_{13} 代表高校重点学科。

2　散点图及数据相关分析

2.1　散点图

散点图能够直观展示两变量之间关系的强弱程度,反映变量间的总体关系趋势。为研究高校生源质量指数与各影响因素之间的相关关系,本文依次绘制出高校生源质量指数与各影响因素之间的散点图,此处仅以高校生源质量指数与院士数(见图 1)、综合声誉指数(见图 2)之间的散点图为例。

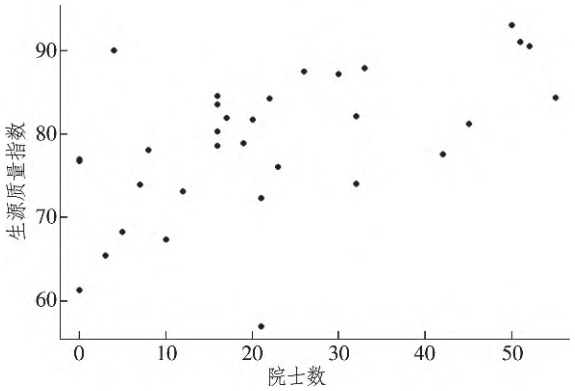

图 1　高校生源质量指数与院士数之间的散点图

Fig.1 Scatter diagram between the quality index of college students and the number of academicians

如图 1 所示,高校生源质量指数与院士数之间存在一定线性趋势,表明两者之间可能存在线性相关关系。

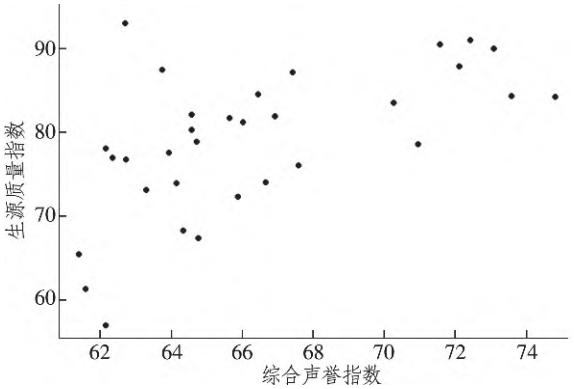

图 2　高校生源质量指数与综合声誉指数之间的散点图

Fig.2 Scatter diagram between the quality index of college students and comprehensive reputation index

同理,由图 2 可知,高校生源质量指数和综合声誉指数之间存在一定线性趋势。

2.2　数据相关性分析

相关分析是研究两个或两个以上处于同等地位的随机变量间的相关关系的统计分析方法。分析高校生源质量指数与院士数、综合声誉指数之间的线性相关系数[7]和秩相关系数,结果见表 1。

表 1　高校生源质量指数与院士数和综合声誉指数的线性相关性与秩相关检验结果

Tab.1 Linear correlation and rank correlation test results of college student quality index, number of academicians and comprehensive reputation index

影响因素	线性相关		秩相关	
	相关系数 r_1	P 值	相关系数 r_2	P 值
院士数	0.562	<0.001	0.551	<0.001
综合声誉指数	0.588	<0.001	0.584	<0.001

高校生源质量指数与院士数之间的线性相关系数 $r_1 = 0.562$,相应的 P 值小于 0.05,拒绝原假设,存在统计学差异,说明高校生源质量指数与院士数呈现低度线性相关关系。秩相关系数 $r_2 = 0.551 > 0$,相应的 P 值小于 0.05,拒绝原假设,存在统计学差异,说明生源质量指数与影响因素院士数低度相关且为正相关关系。

同理,高校生源质量指数与综合声誉指数之间的线性相关系数 $r_1 = 0.588$,相应的 P 值小于 0.05,拒绝原假设,存在统计学差异,说明高校生

源质量指数与综合声誉指数呈现低度线性相关关系。秩相关系数 $r_2 = 0.584 > 0$,相应的 P 值小于 0.05,拒绝原假设,存在统计学差异,说明生源质量指数与影响因素综合声誉指数低度相关且为正相关关系。

3 多元线性回归方法及分析

3.1 多元线性回归分析模型

多元线性回归的基本原理是利用最小二乘法对多个自变量之间的关系进行建模。多元线性回归模型的一般形式:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

其中,Y 表示高校生源质量指数,$X_i(i = 1,2,\cdots,k)$ 代表相关指标,k 为指标数量,β_0 代表回归常数,β_i 代表回归系数,ε 表示随机误差项。

多元线性回归模型的基本假设:

1) 随机误差项的期望值为 0,即 $E(\varepsilon) = 0$;

2) 对于解释变量的所有观测值,随机误差项有相同的方差,即 $Var(\varepsilon) = \sigma^2$;

3) 随机误差项之间彼此不相关;

4) 解释变量不是随机变量而是确定性变量,与随机误差项之间相互独立;

5) 解释变量彼此之间不存在精确的(完全的)线性关系,即解释变量的样本观测值矩阵是满秩矩阵;

6) 随机误差项服从正态分布。

在多元回归中有多个解释变量,需要有足够的证据证明所有变量联合起来对被解释变量有显著的影响,所以有必要进行 F 检验[8]。

F 检验的原假设 H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$;

F 检验的备择假设 H_1: $\beta_j(j = 1,2,\cdots,k)$ 不全为 0;

F 统计量:

$$F = \frac{S_{ESS}/(k-1)}{S_{RSS}/(n-k)} \sim F(k-1, n-k),$$

其中,S_{ESS} 表示回归平方和,S_{RSS} 表示残差平方和,n 代表样本容量,k 代表解释变量的个数。

多元线性回归中,回归方程的显著不代表所有自变量对 Y 的影响都显著,故而希望从回归方程中删除次要的变量,建立简单的回归方程,这就需要对每个自变量做显著性检验(t 检验)[8]。

H_0: $\beta_j = 0, j = 1,2,\cdots,k$;

H_1: $\beta_j \neq 0, j = 1,2,\cdots,k$;

统计量:

$$t = \frac{\hat{\beta}_j - \beta_j}{E_s(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k),$$

其中:$\hat{\beta}$ 表示参数 β 的估计值,通过最小二乘估计得到,其值为 $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$;$E_s(\hat{\beta}_j)$ 表示 $\hat{\beta}_j$ 的标准误差,$\hat{\sigma}$ 表示回归标准差,$c_{jj} = (X'X)^{-1}$,X 是一个 $n \times (p+1)$ 阶的设计矩阵。

3.2 数据分析及处理和模型建立

本文将高校生源质量指数作为因变量,记作 Y,将理工类、科学研究指数等影响因素作为自变量,对所有变量进行回归分析,建立回归模型,得到各变量的回归系数(见表 2);并对所有解释变量使用方差扩大因子法进行共线性诊断,观察变量之间是否存在多重共线性。

表 2 各变量回归系数表

Tab.2 Regression coefficients of each variable

变量名称	系数	t 值	P 值	容忍度	VIF 值
(常数项)		1.643	0.119		
X_1	0.512	3.467	0.003	0.347	2.883
X_2	0.230	1.714	0.105	0.421	2.377
X_3	0.189	1.493	0.154	0.471	2.122
X_4	−0.108	−0.570	0.576	0.212	4.718
X_5	0.028	0.254	0.803	0.640	1.563
X_6	−0.029	−0.201	0.843	0.361	2.774
X_7	0.137	0.618	0.545	0.153	6.516
X_8	−0.382	−2.375	0.030	0.292	3.419
X_9	−0.104	−1.009	0.327	0.712	1.405
X_{10}	0.358	1.728	0.102	0.176	5.671
X_{11}	0.174	0.829	0.418	0.171	5.850
X_{12}	0.065	0.539	0.597	0.524	1.907
X_{13}	0.397	1.607	0.126	0.124	8.079

方差膨胀因子 VIF 值不大于 10 表明不存在强多重共线性,否则存在强多重共线性[9]。由表 2 可知,所有方差膨胀因子均小于 10,表明此回归方程不存在强多重共线性。影响因素中只有理工类和本科生人数的 P 值小于 0.05,即只有这两个因素通过了显著性检验,这与设想的结果不同。为了更加深入地研究高校生源质量的影响因素,本文决定采用逐步回归法建立模型进行研究。

逐步回归分析的基本思想是通过逐个引进变量,对每次引进的解释变量都进行检验,如果检验结果表明后引进的解释变量使得原先引进的解释

变量变得不再显著时,就要将其删除,这样就可保证每次在引进新的变量之前,回归方程中只含有显著性变量[10—11]。如此反复,以得到最优的解释变量集。

将所有自变量导入,使用逐步回归法,得到模型检验结果,见表 3。

表 3　模型检验结果

Tab.3 Test results of the models

模型	影响因素	系数	t 值	P 值
1	常数项		−0.357	0.724
	综合声誉指数	0.588	3.977	0.000
2	常数项		0.380	0.707
	综合声誉指数	0.551	4.935	0.000
	本科生人数	−0.545	−4.879	0.000
3	常数项		−0.762	0.452
	综合声誉指数	0.650	6.509	0.000
	本科生人数	−0.438	−4.360	0.000
	理工类	0.353	3.403	0.002
4	常数项		1.703	0.100
	综合声誉指数	0.305	2.085	0.047
	本科生人数	−0.578	−5.734	0.000
	理工类	0.315	3.395	0.002
	重点学科	0.431	2.952	0.006

由表 3 可知,最终仅保留了 4 个影响因素,分别为综合声誉指数、本科生人数、是否是理工类院校以及重点学科。因为常数项 P 值为 0.100,大于显著性水平 0.05,没有通过检验,说明常数项对于生源质量指数的影响不存在统计学差异,所以尝试去掉常量再次进行逐步回归,结果见表 4。

表 4　去除常数项后模型检验结果

Tab.4 Test results of the models after removing the constant term

模型	影响因素	系数	t 值	P 值
1	综合声誉指数	0.996	63.516	0.000
2	综合声誉指数	1.104	44.566	0.000
	本科生人数	−0.123	−4.958	0.000
3	综合声誉指数	1.070	45.341	0.000
	本科生人数	−0.105	−4.760	0.000
	理工类	0.040	3.370	0.002
4	综合声誉指数	1.031	38.477	0.000
	本科生人数	−0.112	−5.469	0.000
	理工类	0.045	4.048	0.000
	重点学科	0.049	2.477	0.020

由表 4 可知,去除常数项后,所有变量的 P 值均小于显著性水平 0.05,均通过 t 检验,说明这些变量对于生源质量的影响存在统计学差异。观察各个变量的系数可以看出: 综合声誉指数、理工类和重点学科 3 个变量的系数为正,说明这 3 个变量对高校生源质量具有显著的正向影响,即这 3 个变量值越大,高校生源质量越好; 本科生人数的系数为负,说明该变量对高校生源质量具有显著的负向影响,即变量值越大,高校生源质量越差。

对最后确定的最优模型进行模型评价,见表 5。

表 5　模型评价结果

Tab.5 Evaluation results of the models

模型	R	R^2	修正的 R^2	DW 值
1	0.588	0.345	0.323	
2	0.800	0.640	0.616	
3	0.863	0.746	0.718	
4	0.899	0.808	0.779	2.014

由表 5 可知,R^2 为 0.808,调整后的 R^2 为 0.779,说明该模型的拟合效果较好[12],且 DW 值为 2.014,接近临界值 2,表明有很大把握认为该模型变量不存在自相关性[13]。

4　Lasso 回归及结果分析

4.1　Lasso 回归基本原理

Lasso 回归方法通过构造一个惩罚函数得到一个较为精炼的模型,达到压缩回归系数的目的,是一种处理具有复共线性数据的有偏估计[14]。

岭回归无法降低模型复杂度,而 Lasso 回归是在岭回归基础上的优化[15],可以直接将系数惩罚压缩至零,达到降低模型复杂度的目的。Lasso 回归的目标函数可以表示为:

$$L(b) = \sum (y - Xb)^2 + \lambda \| b \|_1 = \sum (y - Xb)^2 + \sum \lambda |b|,$$

其中: $\lambda \| b \|_1$ 为函数的惩罚项,λ 为惩罚系数; $\| b \|_1$ 为回归系数 b 的正则,表示所有回归系数绝对值的和。

4.2　结果分析

安装 glmnet 包,使用 Lasso 回归对多个影响因素进行变量筛选[16],计算各影响因素回归系数,结果见表 6。

表6 筛选处理后的各影响因素回归系数

Tab.6 Regression coefficient and exponential coefficient of each influence factor after sorting

变量名称	回归系数	Exp(回归系数)
本科生人数	−0.000 307 513	9.996 925e−01
科学研究指数	0.025 536 404	1.025 865e+00
社会服务指数	0.059 503 101	1.061 309e+00
院士数	0.062 517 241	1.064 513e+00
重点学科	0.194 989 389	1.215 298e+00
综合声誉指数	0.551 877 491	1.736 510e+00
北京上海	2.008 777 109	7.454 196e+00
理工类	4.146 539 546	6.321 487e+01
截距项	37.392 269 931	1.734 828e+16

从13个自变量中筛选得到8个符合要求的自变量,分别是:本科生人数、科学研究、社会服务指数、院士数、重点学科、综合声誉指数、北京上海和理工类,对应的回归系数见表6。另外5个变量为:综合类、一线城市、沿海城市、本科就业率和师生比,它们对应的系数值为0,说明这5个变量对高校生源质量指数没有显著意义。故可以将Lasso回归模型表达为:

$$Y = 37.392 + 4.167X_1 + 0.026X_3 + 0.063X_6 + 0.552X_7 - 3.075 \times 10^{-4}X_8 + 2.009X_{10} + 0.060X_{11} + + 0.195X_{13}。$$

对经过筛选处理的变量参数进行检验,结果见表7。

表7 各影响因素的检验结果

Tab.7 Test results of each influencing factor

变量名称	参数估计值	标准误差	t值	P值
截距项	3.387e+01	2.252e+01	1.504	0.146
本科生人数	−3.272e−04	9.652e−05	−3.390	0.003
科学研究指数	6.734e−02	6.676e−02	1.009	0.324
社会服务指数	1.420e−01	1.783e−01	0.796	0.434
院士数	4.943e−02	7.010e−02	0.705	0.488
重点学科	2.554e−01	1.877e−01	1.360	0.187
综合声誉指数	4.794e−01	4.179e−01	1.147	0.263
北京上海	2.939e+00	2.016e+00	1.458	0.158
理工类	6.452e+00	2.258e+00	2.858	0.009

由表7可知,8个符合要求的自变量中只有本科生人数和理工类这两个变量的P值小于0.05,通过了t检验,表明这两个变量对于生源质量的影响存在统计学差异[17]。该结果与预期不一致,剔除掉最不显著的变量,即院士数,再次进行检验,结果见表8。

表8 剔除院士数后各影响因素的检验结果

Tab.8 Test results of influencing factors after excluding the number of academicians

变量名称	参数估计值	标准误差	t值	P值
截距项	3.598e+01	2.208e+01	1.629	0.116 34
本科生人数	−3.428e−04	9.298e−05	−3.686	0.001 16
科学研究指数	6.041e−02	6.533e−02	0.925	0.364 41
社会服务指数	1.154e−01	1.724e−01	0.669	0.509 75
重点学科	3.339e−01	1.494e−01	2.235	0.034 99
综合声誉指数	4.790e−01	4.135e−01	1.158	0.258 08
北京上海	3.055e+00	1.988e+00	1.537	0.137 46
理工类	6.918e+00	2.136e+00	3.239	0.003 49

由表8可知,将变量院士数剔除之后,除了本科生人数和理工类以外,重点学科这一变量也通过了检验。重复上面方法,继续剔除变量社会服务指数,进行检验,结果见表9。

表9 剔除社会服务指数后各影响因素的检验结果

Tab.9 Test results of influencing factors after excluding social service index

变量名称	参数估计值	标准误差	t值	P值
截距项	3.370e+01	2.158e+01	1.562	0.131
本科生人数	−3.510e−04	9.115e−05	−3.851	0.001
科学研究指数	5.208e−02	6.343e−02	0.821	0.419
重点学科	3.553e−01	1.444e−01	2.462	0.021
综合声誉指数	6.390e−01	3.337e−01	1.915	0.067
北京上海	3.074e+00	1.966e+00	1.563	0.131
理工类	6.541e+00	2.037e+00	3.211	0.004

由表9可知,将社会服务指数这个变量剔除之后,仍然只有本科生人数、理工类和重点学科这3个变量通过检验。

按照上述方法依次将变量科学研究指数和北京上海剔除,得到最优模型,其中包含的影响因素为:本科生人数、重点学科、综合声誉指数、理工类。各影响因素的检验结果见表10。

表10 最优模型中各影响因素的检验结果

Tab.10 Test results of the influencing factors in the optimal model

变量名称	参数估计值	标准误差	t值	P值
截距项	3.572e+01	2.097e+01	1.703	0.100
本科生人数	−4.436e−04	7.736e−05	−5.734	4.28e−06
重点学科	3.654e−01	1.238e−01	2.952	0.006
综合声誉指数	6.794e−01	3.258e−01	2.085	0.047
理工类	6.524e+00	1.922e+00	3.395	0.002

由表10可知,将其他4个变量一一剔除后,所剩变量本科生人数、重点学科、综合声誉指数和理工类的P值均小于0.05,通过了t检验,表明这4个变量对于生源质量的影响存在统计学差异。这与使用逐步回归法得到的结果一致,增加了本

文研究所得的高校生源质量影响因素的可信度，使研究结果更加具有说服力。

5　结语

本文分别运用逐步回归和 Lasso 回归进行分析，得出以下结论：影响高校生源质量指数的因素主要有本科生人数、重点学科、综合声誉指数以及高校是否属于理工类学校。同时结合相关性分析结果得出：重点学科数、综合声誉指数以及理工类与生源质量指数呈线性正相关，本科生人数与生源质量指数呈线性负相关。进一步得到推论：增加重点学科数、提高综合声誉指数以及降低本科生人数均可以提高生源质量；由于是否是理工类院校不属于可控因素，故本文不针对该影响因素提出具体的措施。

根据上述分析结果，本文认为各高校采取以下措施，可能对提高高校生源质量有所帮助：

1）适当降低本科生的招生人数；

2）增加重点学科的数量；

3）从同行声誉，媒体声誉和国际声誉三方面入手提高高校综合声誉。

本文的局限性在于只选取了 32 所 985、211 高校的相关数据进行研究，研究范围较小，在后续研究中考虑将研究范围扩展至普通双非高校，以使研究得出的结果更具有参考性和普适性。

参考文献：

[1] 邵风侠.影响高校生源质量提高的因素分析及对策建议[J].北京教育(高教),2019(12):65-67.

[2] 史秋衡，矫怡程.不同类型高校本科生源质量的实证研究：基于“国家大学生学习情况调查”的数据分析[J].复旦教育论坛,2014,12(1):18-23.

[3] 赵良君，申静.高校生源质量影响因素分析及对策研究[J].继续教育研究,2009(7):127-128.

[4] 侯爽，李磊.高校生源质量控制和就业形势分析[J].科教文汇(中旬刊),2018(5):142-143.

[5] 赵飞，张锦宗，朱瑜馨.“985”高校本科生生源质量评价及影响因素分析[J].当代教育论坛,2015(6):6-13.

[6] 肖维.高校生源质量影响因素分析[D].大连：大连理工大学,2016.

[7] 李秀敏，江卫华.相关系数与相关性度量[J].数学的实践与认识,2006,36(12):188-192.

[8] 刘明，王仁曾.基于 t 检验的逐步回归的改进[J].统计与决策,2012(6):16-19.

[9] 马雨阳，官海翔，杨昊轩，等.利用地形、土壤和作物信息辅助提高东北漫岗地数字高程模型精度的新方法[J].中国农业科学,2021,54(8):1715-1727.

[10] 游士兵，严研.逐步回归分析法及其应用[J].统计与决策,2017(14):31-35.

[11] 李楠，焦庆宇，樊瑞，等.基于逐步回归法航空器滑行时间影响因素研究[J].计算机仿真,2021,38(9):57-63.

[12] 王巧英.回归估计标准误差与可决系数的比较[J].统计与决策,2006(23):141.

[13] 赵卫亚.DW 检验的局限性与模型的高阶自相关检验[J].统计与决策,2004(1):18-19.

[14] 方匡南，章贵军，张惠颖.基于 Lasso-logistic 模型的个人信用风险预警方法[J].数量经济技术经济研究,2014,31(2):125-136.

[15] 朱海龙，李萍萍.基于岭回归和 LASSO 回归的安徽省财政收入影响因素分析[J].江西理工大学学报,2022,43(1):59-65.

[16] 赵俊琴，王彤，王慧，等.Lasso-惩罚计分检验在小样本回归模型自变量筛选与统计推断中的应用[J].中华疾病控制杂志,2015,19(5):507-509.

[17] 李阳，陈晓泓，王一梅，等.基于 LASSO 变量选择联合贝叶斯网络构建恶性肿瘤相关急性肾损伤(AKI)风险预测模型[J].复旦学报(医学版),2020,47(4):521-530.

Influencing Factors of College Student Quality Based on Multiple Linear Regression and Lasso Regression

LI Xueyang, SHAO Xigao

(School of Mathematics and Statistical Science, Ludong University, Yantai 264039, China)

Abstract: In order to further improve the quality of college students, several factors that may affect the quality of college students were selected in this paper, and the required data was obtained from the authoritative data website for research and analysis. Firstly, the correlation between the quality of college students and various fac-

tors was analyzed,and a multiple linear regression model was established by using SPSS software.Secondly,the model was tested for significance,and the stepwise regression method was used to select variables,and finally the optimal multiple linear regression model was obtained.Based on the above data,Lasso regression was used to select variables to test the significance of the variables after screening,and finally determine the factors that have a significant impact on the quality of college students.The results obtained by the two methods in this paper show that the number of undergraduates,key disciplines,comprehensive reputation index and whether a university is a science and engineering school have significant impacts on the quality of college students.

Keywords: multiple linear regression; Lasso regression; significance test; correlation

(责任编辑 李秀芳)

(上接第328页)

Abstract ID: 1673-8020(2022) 04-0320-EA

Spatial Distribution and Tourism Response of Cultural Heritage in Shandong Province: Taking the Cultural Heritage Protection Unit as an Example

XU Huimin[a], WANG Jialing[a], JIN Shizhu[a,b]

(a.School of Geography and Marine Science; b.College of Integration Science, Yanbian University, Yanbian 133000, China)

Abstract: Based on the perspective of geography, Excel, ArcGIS and SPSS software were used in this paper to quantitatively and spatially analyze 226 national and 1711 provincial cultural heritage protection units in Shandong Province.The results are as follows: 1) There is a large gap between the number of national and provincial cultural heritage protection units in Shandong Province, and the differences in the number of cultural heritage protection units in different batches are obvious; the structures of the type of the national and provincial cultural heritage protection units in Shandong Province are shown to have similarities, and they are mainly composed of ancient ruins, ancient buildings and important historical sites and representative buildings in modern times, the proportion of other cultural heritage protection units is low; the national cultural heritage protection unit and provincial cultural heritage protection unit are shown to be clustered in space, with polar core area, high density area, sub-density area, intensive area and banded area as the distribution pattern; 2) The distribution of cultural heritage protection unit in Shandong Province is affected by many factors such as history and culture, natural environment and human environment; 3) To a certain extent, the tourism industry in areas is supported by the material cultural heritage resources of Shandong Province, but foreign exchange earnings from tourism cannot be explained by its distribution, which is related to the economic development level of the destination, type of tourism resources and tourist preferences and other factors.

Keywords: cultural heritage protection unit; spatial distribution characteristics; the causes of formation; tourism response; Shandong Province

(责任编辑 李秀芳)