Statistics and Application 统计学与应用, 2023, 12(2), 381-390
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122041

Hans汉斯

我国教育水平影响因素研究

万家豪

北方工业大学理学院，北京

收稿日期：2023年3月24日；录用日期：2023年4月14日；发布日期：2023年4月26日

摘　要

以2015年到2019年31个省份的人均受教育年限为研究对象，从人口、政策、经济、科技水平、服务供给五个大方面选择了10个影响因素。首先运用弹性网对影响教育水平的因素进行筛选压缩，并选用岭回归模型、Lasso回归模型作为对比。最终得出较为合适的变量进行参数估计，最后对模型的准确率进行预测。得出最终影响因素分析结果。

关键词

教育水平，弹性网络，线性回归，Lasso

Research on Influencing Factors of China's Educational Level

Jiahao Wan

College of Science, North China University of Technology, Beijing

Received: Mar. 24th, 2023; accepted: Apr. 14th, 2023; published: Apr. 26th, 2023

Abstract

Taking the per capita years of schooling in 31 provinces from 2015 to 2019 as the research object, 10 influencing factors were selected from five aspects: population, policy, economy, scientific and technological level, and service supply. Firstly, elastic net was used to select and compress the factors affecting education level, and Ridge regression model and Lasso regression model were selected as comparison. Finally, more appropriate variables are obtained for parameter estimation, and finally the accuracy of the model is predicted. The final analysis results of influencing factors are obtained.

文章引用: 万家豪. 我国教育水平影响因素研究[J]. 统计学与应用, 2023, 12(2): 381-390.
DOI: 10.12677/sa.2023.122041

Keywords

Education Level, Elastic Networks, Linear Regression, Lasso

1. 引言

改革开放以来，我国人均受教育水平不断提高，中国人力资本储备也由此增加。随着科教兴国、人才强国战略以及义务教育政策的实施，使得越来越多的人更容易接收到教育。但我国教育水平仍然是世界的较低水平。在现阶段发展对人才的竞争越来越激烈的情况下，如果促进我国教育水平的提高就变得尤为重要。

教育水平的增长对于经济的发展和科技进步有明显的促进作用这一点是不争的事实。也正是因为我国正在稳步发展经济和科技，因此对于提高国民教育水平来说是必要的。然而根据我国国情来看，人口素质依旧偏低，而且国土面积虽然很大，但西部地区地势较高、山林交错的特点导致交通不便、人群居住较散等问题也成了导致我国教育水平很低的关键因素。

对于教育水平的研究，大部分学者的研究方向均是以教育水平为自变量去探究其对科技发展水平、经济发展水平等的影响。但是很少有研究针对教育水平为因变量，去探究在各种影响因素的情况下，我国各省份教育水平的情况。本文正是再次背景下运用机器学习回归中一些方法来进行相关研究。

2. 影响因素分析

教育作为一个大众化的服务行业，收到很多因素的直接或间接影响。为了对受教育水平有更有效的分析，结合参考相关文献[1]和国家目前经济社会发展的实际情况。从人口、政策、经济、科技水平、服务供给、个人平均水平六大方面进行变量选取。针对以上方面构建概念模型，并分别解释各指标对教育水平的影响作用机理。从而更好的确认选取变量的角度以及是否有充分解释效果。如图 1 所示。

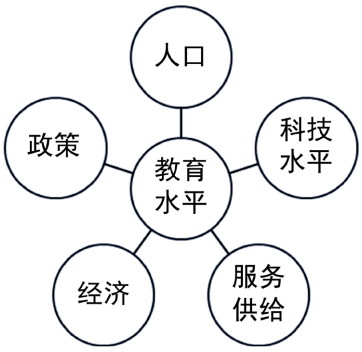

Figure 1. Conceptual model
图 1. 概念模型

在人口方面，教育水平的直接体现就是人均受教育年限。在一系列对于人口因素中，男女比例无疑会对教育产生一定影响。部分农村及信息不发达地区依旧有重男轻女的腐朽思想，因此对男生进行教育，而女生则放弃教育。由此出现教育水平的区别，由此不仅说明了男女比例对教育的影响，同时也说明了

城乡居住地的影响。城乡直接交通运输的发达程度存在差异，导致了人们接触外界的眼界存在不同。一系列差异最终使城市更倾向于教育投入，而农村部分则更加关注与目前生活温饱情况，对教育忽视度较高。然而也有一部分情况显示，城市人接触到外界之后变得更加放纵从而放弃教育，而农村为了能更好走出去发展而选择加大教育力度。此外人口的自然增长也同样影响着教育水平的变化。而这些影响整体上是正向促进了教育水平的提高还是负向限制了教育的发展，则需要具体数据来进行定量分析。

在经济方面，随着城镇化水平的推进和居民收入的普遍提高，一方面存在人们在满足自身衣食住行的需求之后在教育等方面加大投入，另外一方面同样有人更加注重奢侈的生活而放弃教育的机会。

在政策方面，国家始终在推行科教兴国和人才强国和战略、义务教育、高校扩招政策、经费投入以及教育公平。在诸多大环境的影响下，我国教育水平理应向更好的方向转变。然而根据获得数据显示并没有明显优势变化，则说明还有更大的负向影响。

在科技水平方面，科技的创兴和教育水平的提高往往是双向促进的。随着科技的不断发展，教育所能使用的资源和先进设备越来越多，从而推进这教育的发展。

在服务供给方面，学校数量的增加和专职教师数的增加使得可以同时接受更多人进行教育培养。学校数量的不断增加，各个地区政府针对教育方面的政策以及各种社会公益事业对于教育的重视，使得学校的创办更加简单快捷。而随着教师工资待遇的提升，国家专职教师数也出现了明显的增多。对于教育培养有很大帮助。

综上所述，从每一个方面可能都存在积极或者消极作用，因此需要定量进行分析。

3. 模型介绍

从以上情况可以看出，影响我国教育水平的因素很多，而且涉及到很多方面。因此变量选择是否全面和精确是重要的，本节这分别介绍岭回归、Lasso 以及弹性网络的变量选择方法对所给数据处理的优势。

3.1. 岭回归

回归模型的标准化形式为：$\tilde{Y}=\theta_1\tilde{X}_1+\cdots+\theta_P\tilde{X}_P+\varepsilon'$。

岭回归是线性回归的正则化[2]：将等于 $\alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2$ 的正则化项添加到成本函数中。这使得学习算法不仅要拟合数据，而且还是模型权重尽可能小。超参数 α 控制要对模型进行正则化程度。岭回归的成本函数：

$$J(\theta)=\mathrm{MSE}(\theta)+\alpha\frac{1}{2}\sum_{i=1}^{n}\theta_i^2 \text{。}$$

岭回归的目标是生成一个系数稳定的回归方程。系数稳定和选择好的变量有相似的目标，所以可以使用岭回归来进行变量选择。

条件：1) 剔除系数稳定但绝对值很小的变量。因为岭回归处理的是标准化数据，因此不同系数的数值大小可直接比较。2) 剔除系数不稳定而无预测能力的变量。即趋于 0 的不稳定系数。3) 剔除一个或多个系数不稳定的变量。用剩下的 p 个变量建立回归方程。

以上每进行一步就用剩下变量重新拟合模型，然后在进行下一步。

3.2. Lasso 回归

Lasso 回归是线性回归的另外一种正则化，叫做最小绝对收敛和选择算子回归[3]。其本质是用模型系数的绝对值函数作为惩罚因子的最小二乘估计，通过把 OLS 方法估计得到的系数压缩为 0，从而实现变量筛选[4]。

成本函数为：$J(\theta)=\mathrm{MSE}(\theta)+\alpha\sum_{i=1}^{n}|\theta|$，后项增加项为权重向量的 l_1 范数。

Lasso 回归一个重要特点就是它倾向于完全消除掉最不重要的特征的权重。Lasso 回归有传统方法的优点，并且有更有效的算法：最小角回归算法[5]。以此达到系数收缩和变量选择的目的。

局限：1) 当预测变量个数大于样本数时，最多只能选择样本数个变量；2) 如果有一组变量两两相关性很高，则只能选取其中一个；3) 当样本数大于预测变量数时，如果变量之间高度相关，则 Lasso 的预测水平并不高[6]。

3.3. 弹性网络

弹性网络介于岭回归和 Lasso 回归的中间地带，是岭回归和 Lasso 回归的正则化简单混合。成本函数为[7]：

$$J(\theta)=\mathrm{MSE}(\theta)+\gamma\alpha\sum_{i}^{n}|\theta|+\alpha\frac{1-\gamma}{2}\sum_{i=1}^{n}\theta_i^2$$

对于弹性网络算法，只要通过变化将弹性网络方法的解表达成类似于 Lasso 方法的节，就能利用 Lars 算法得出弹性网络方法的节。弹性网络回归在自动选择变量的同时还能实现连续收缩，并且具有很好的群组效应，预测能力上更优与 Lasso 方法[8]。

4. 实证分析

4.1. 变量设计

本文重点讨论教育水平的影响因素，参考文献及国内社会实际情况，从人口、政策、科技水平、服务供给、经济五个方面设计了 11 个预测变量。如表 1 所示。

Table 1. Influence factor
表 1. 影响因素

一级指标	二级指标	变量
教育水平	受教育平均年限	y
人口	男女比例(女 = 1)	NNBL
	人口增长率	RKZZL
	城乡比例(乡 = 1)	CXBL
国家政策	教育经费(万元)	GJZC
科技水平	R&D 经费(万元)	KJSP
经济	人均可支配收入(万元)	RJKZPSR
	人均教育消费占比(%)	RJJYXFZB
服务供给	生师比(师 = 1)	SSB
	学校数(每十万人)	XXS
	教学质量	JXZL
	专职教师数(每十万人)	ZZJSS

注：1) 生师比为统计年鉴中直接数据；2) 学校数为小学以上地区学校总数/地区年末常住人口数；3) 教学质量设计为：当年毕业人数/招生人数。

4.2. 数据来源

文章所有原始数据均来自国家统计局网站上下载，包括 2015 年到 2019 年五年内 31 个省的数据供

155 例。

文章采用机器学习方法随机抽取 70%样本作为测试集，30%为训练集。表 2 为样本数据详情。

Table 2. Sample
表 2. 样本详情

样本	样本数
测试集	116
训练集	39

5. 模型估计结果

文章选取了 11 个变量对教育水平进行分析，但所选取的变量并非都对其有显著影响，且变量之间可能存在多重共线性等问题。因此首先用线性回归做一下多重共线性的检测已经变量是否显著的初步观察。

5.1. 线性回归

根据图 2 简单线性回归结果，存在五个预测变量的 t-检验的 p 值较大。说明不够显著。分别为：NNBL，KJSP，RJJYXFZB，SSB，ZZJSS。通过计算各个变量的相关系数矩阵特征值已经方差膨胀因子(VIF)，如表 3。

```
==============================================================================
Dep. Variable:                   PJNX   R-squared:                       0.723
Model:                            OLS   Adj. R-squared:                  0.702
Method:                 Least Squares   F-statistic:                     33.91
Date:                Tue, 15 Dec 2020   Prob (F-statistic):           1.68e-34
Time:                        11:57:53   Log-Likelihood:                -139.69
No. Observations:                 155   AIC:                             303.4
Df Residuals:                     143   BIC:                             339.9
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
constant       4.1805      1.815      2.304      0.023       0.594       7.767
NNBL           0.0131      0.015      0.897      0.371      -0.016       0.042
RKZZL         -0.0946      0.023     -4.035      0.000      -0.141      -0.048
CXBL           0.4018      0.101      3.989      0.000       0.203       0.601
GJZC       -4.033e-08   1.81e-08     -2.223      0.028   -7.62e-08   -4.47e-09
KJSP          -0.0044      0.003     -1.555      0.122      -0.010       0.001
RJKZPSR     3.769e-05   1.55e-05      2.438      0.016    7.13e-06    6.82e-05
RJJYXFZB       0.0027      0.012      0.234      0.816      -0.020       0.026
SSB            0.0552      0.047      1.170      0.244      -0.038       0.148
XXS            0.1435      0.065      2.201      0.029       0.015       0.272
JXZL           0.0293      0.009      3.394      0.001       0.012       0.046
ZZJSS          0.0698      0.056      1.244      0.215      -0.041       0.181
==============================================================================
Omnibus:                       23.097   Durbin-Watson:                   1.576
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.933
Skew:                          -0.763   Prob(JB):                     9.55e-09
Kurtosis:                       4.840   Cond. No.                     5.36e+08
==============================================================================
```

Figure 2. Results of linear regression
图 2. 线性回归结果

Table 3. Eigenvalue and VIF
表 3. 特征值及方差膨胀因子

变量	相关系数	VIF
常数项		1326.137
NNBL	3.703	1.561

Continued

RKZZL	1.961	1.954
CXBL	1.245	9.532
GJZC	1.106	8.158
KJSP	1.028	1.112
RJKZPSR	0.041	11.569
RJJYXFZB	0.099	1.152
SSB	0.266	2.014
XXS	0.358	1.901
JXZL	0.672	6.579
ZZJSS	0.521	2.485

从表上看，存在两个较小的特征值 0.099 和 0.041，以及一个大于 10 的方差膨胀因子。由此可以说明选择的预测变量之间是存在多重共线性的。则接下来就需要进行变量的筛选。

5.2. 岭回归筛选

根据第二节所示方法，对原始数据首先进行标准化处理，克服量纲影响。中心标准化即使：

$$\frac{1}{n}\sum_i y_i = 0, \frac{1}{n}\sum_i x_{ij} = 0, \frac{1}{n}\sum_i x_{ij}^2 = 1$$

对原始标准化之后的数据选择 33 个 0~1 之间的 α 值使用 python 中 sklearn 库中的岭回归方法，获得了随 α 之变化的各变量系数。如下图 3 所示。其中系数绝对值小于 0.1 的当做较小影响因素进行删除，第一次结果显示：NNBL，RJJYXFZB，SSB 三个变量被删除；之后再使用删除过数据的新样本数据进行第二次岭回归拟合，如图 3 右所示，同理显示 ZZJSS 被删除；接下来继续按照此方法对之后数据再次拟合，结果如图 4 左所示，同样的删除 KJSP。之后进行第四次岭回归拟合，如图 4 右所示。这次所拟合的结果中系数绝对值均大于 0.1，证明均为显著性变量，由此保留如下变量作为岭回归最后筛选结果。

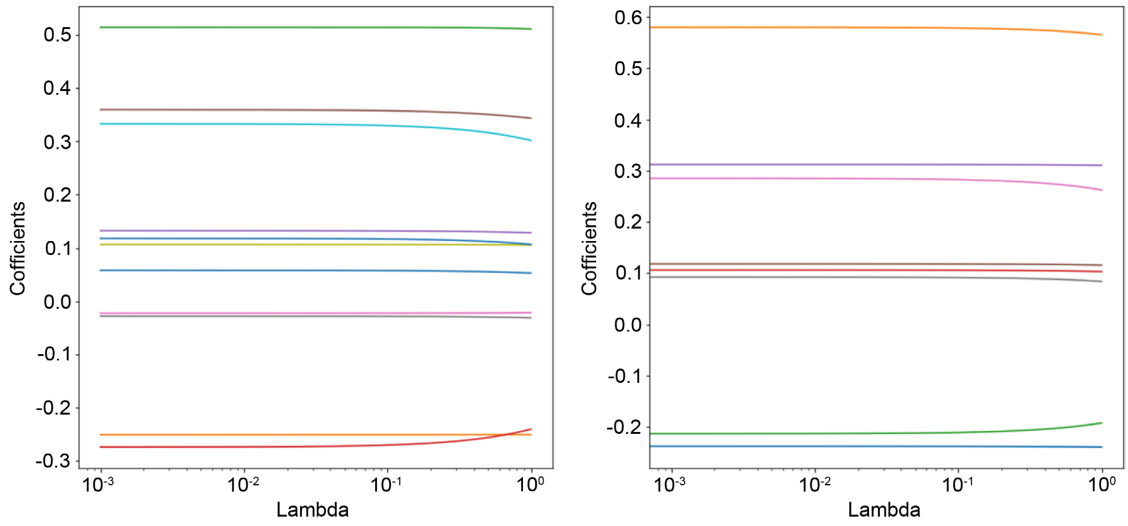

Figure 3. The first and second ridge regression
图 3. 第一次、第二次岭回归

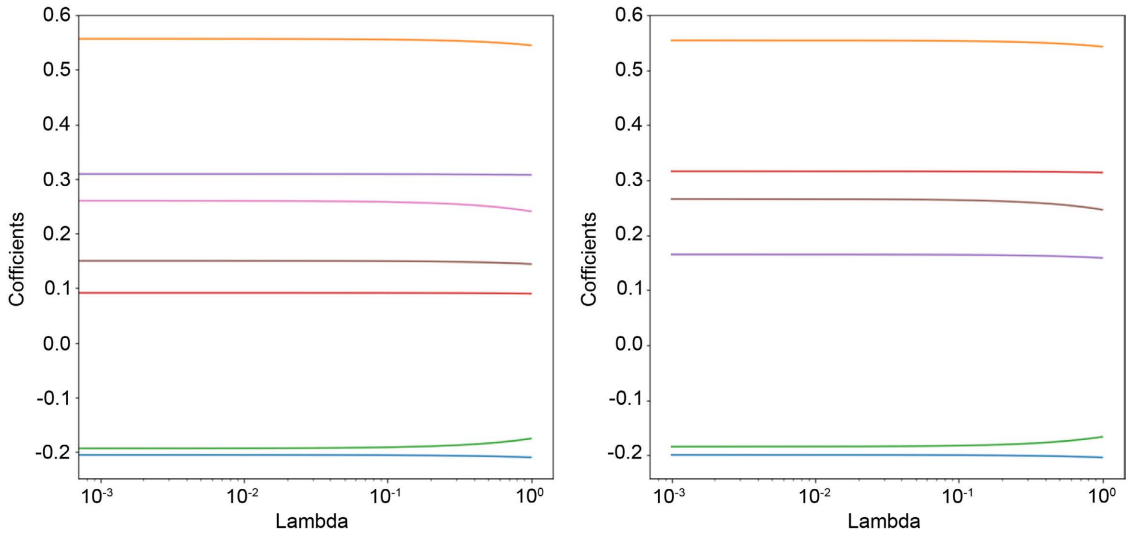

Figure 4. The third and fourth ridge regression
图 4. 第三次、第四次岭回归

综合四次筛选结果，最终保留了六个变量分别为：RKZZL, CXBL, GJZC, RJKZPSR, XXS, JXZL，之后使用 sklearn 库中的交叉验证法获得最优的 α 值为 1.0。因此选择 α 值为 1.0 的岭回归成本函数对剩余的六个变量做岭回归估计，得到其系数如表 4 所示。

Table 4. Results of ridge regression variable filtering
表 4. 岭回归变量筛选结果

变量	Constant	RKZZL	CXBL	GJZC	RJKZPSR	XXS	JXZL
系数	0.022	−0.203	0.545	−0.166	0.311	0.158	0.247

5.3. Lasso 回归

进行 Lasso 回归的第一步同样是对样本进行标准化处理客服量纲影响。之后再 1×10^{-5} 到 1×10^{2} 之间选择 300 个随机数作为拟合 Lasso 回归的 α 值。获得系数关于 α 值得关系如图 5 所示。可知当 α 的值在 1×10^{-3} 到 1×10^{-2} 之间时，对 Lasso 模型的拟合结果系数较多趋近，从而变得稳定。则可初步判断最优的 α 值应该位于 1×10^{-3} 到 1×10^{-2} 之间。

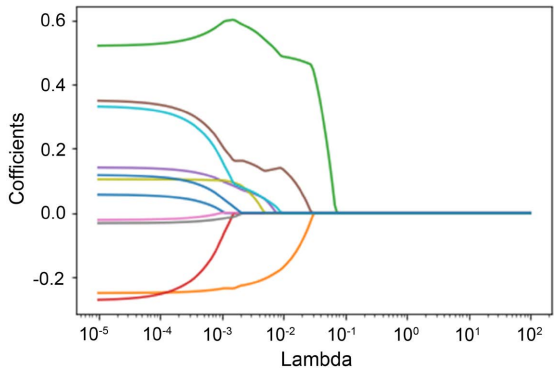

Figure 5. Regression coefficient selection of Lasso
图 5. Lasso 回归系数选择

在此图的基础上使用交叉验证的方法，获得最好的 α 值为 0.007 也刚好在这一区间内。因此选择此 α 值进行模型拟合并获得筛选结果，即提出变量：NNBL, GJZC, RJJYXFZB, SSB, XXS, ZZJSS。如表 5 所示为 Lasso 模型拟合后所选择变量的系数。

Table 5. Results of Lasso regression
表 5. Lasso 回归结果

变量	Constant	RKZZL	CXBL	KJSP	RJKZPSR	JXZL
系数	0.022	−0.203	0.545	−0.166	0.311	0.158

5.4. 弹性网络

依照第二节方法，弹性网络为岭回归和 Lasso 回归的一种简单融合。按照交叉验证的方法最佳 l1 值为 0.5，之后再 1×10^{-5}，1×10^{2} 上选取 100 个 α 的值进行拟合，所得训练集测试集上效果对比如图 6 所示。黑色位置即为本次选择最佳 α 的位置，即 0.115。

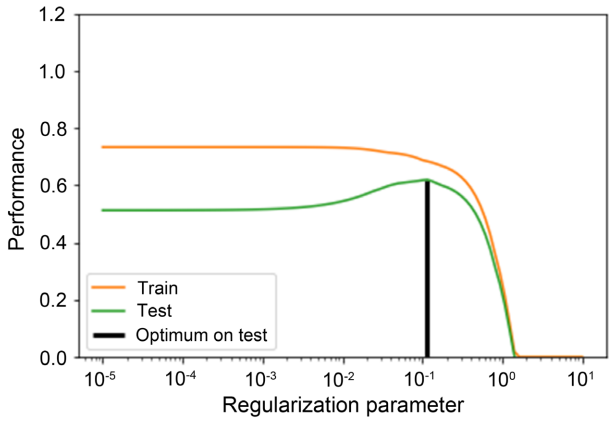

Figure 6. Elastic network filtering
图 6. 弹性网络筛选

使用 $l_1 = 0.5$，$\alpha = 0.115$ 来拟合测试集，最终删除了 NNBL，GJZC，RJJYXFZB，SSB，XXS，ZZJSS。保留下来的变量如表 6 所示。

Table 6. Elastic network final choice
表 6. 弹性网络最后选择

变量	Constant	RKZZL	CXBL	KJSP	RJKZPSR	JXZL
系数	−0.0 (极小)	−0.217	0.431	−0.038	0.192	0.061

6. 模型预测效果评价

为了更准确地选取有效模型，应该选择一些合适的评价指标对以上使用的三组模型最终效果做出评价，获得最优的拟合结果。

6.1. 模型预测误差评价标准

1) 误差均方根 RMSE

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{t=t+1}^{T+n}\left(\hat{y}_t-y_t\right)^2}$$

2) 绝对误差平均 MAE

$$\mathrm{MAE}=\frac{1}{n}\sum_{t=T+1}^{T+n}\left|\hat{y}_t-y_t\right|$$

3) 相对误差绝对值平均 MAPE

$$\mathrm{MAPE}=\frac{1}{n}\sum_{t=T+1}^{T+n}\left|\frac{\hat{y}_t-y_t}{y_t}\right|$$

其中，T 为样本容量、n 表示样本外预测数即测试集样本数、$\hat{y}_t$ 为预测值、y_t 为真实值[9]。

6.2. 预测效果评价

利用三次使用的模型，即岭回归、Lasso 回归、弹性网络。对教育水平测试集进行预测并对预测效果进行评价，结果如表 7 所示。

Table 7. Evaluation criterion
表 7. 评价标准

样本	模型	RMSE	排名	MAE	排名	MAPE	排名
训练集	岭回归	0.508	1	0.377	3	4.325	3
	Lasso 回归	0.540	2	0.369	2	3.431	1
	弹性网络	0.543	3	0.367	1	3.814	2
测试集	岭回归	0.621	1	0.440	2	1.238	3
	Lasso 回归	0.686	3	0.442	3	1.024	1
	弹性网络	0.633	2	0.432	1	1.077	2

根据以上评价标准可以看出，在三个评价标准的选择下，弹性网络在拟合结果的测试上效果相对最好。其次是岭回归模型，最后是 Lasso 模型[10]。

7. 分析和结论

7.1. 分析与启示

1) 对选择的变量方面：城乡比例代表了城市人数与农村人数的差别，而其影响系数为正则说明了城镇人口更加注重于教育，因此国家城镇化发展对教育的提高有促进作用；人均可支配收入的增加使得人们在教育上的投入也逐渐增大，也由此促进了教育的发展；教学质量的提高，毫无疑问促使更多学生接触的教育更高，同时也是其对更高等的知识的探索的提升；人口增长率的增加使得学校在可接纳的人数范围内并不能完全满足所有增加人口的数量。因此导致了很多人口上学比例相对于总人口数变小。间接影响到了作为预测变量的平均受教育年限；科技水平的增加不仅没有促使教育水平的提升，反倒是存在负相关关系。原因可能是由于科技水平的增加，仅在更加发达的城市凸显出来，而教育的水平全国各个地方的总体变现、并且本次选取的科技方面的影响因素仅为 R&D 经费投入金额，并不能完全说明该地区

科技发展水平是否提升。

2) 对删除变量方面：男女比例并没有对教育水平产生任何影响，就说明当今社会更加的公平，人民的思想也更加进步，任何人都有追求更高教育的权利。国家政策代表量中仅含国家经费投入，而国家对教育最大的措施：科教兴国、人才强国战略及义务教育的影响水平并没有很好地体现在变量上，因此被剔除；人均教育消费占比说明和人均可支配收入的影响存在明显相关性关系，被删除；师生比以及学校数则说明，一个地区学校再多、老师再多都并不能展示这个地区教育水平相较于其他地区有多强，而真正影响到的因素应该是该地区教师所展现出的教学水平，即教学质量更加重要。

7.2. 结论

本文选取的数据为 2015 到 2019 年的面板数据作为研究对象，从五个大方面选择了 11 个影响因素，并且运用岭回归、Lasso 回归、弹性网络这三种模型作对比分析，得出以下结论。

1) 模型选择方面：Lasso 回归对模型删减相对于岭回归和弹性网络稍弱，岭回归筛选效果合适，但是存在一个自身判断的问题，即在对最小系数做删除时，对于系数默认为零的范围为自身选择。因此最佳模型为弹性网络。

2) 变量选择方面：a) 城乡比例、人均可支配收入和教学质量与教育水平呈正相关关系，其中系数最大的为城乡比例；b) 人口增长率和科技水平对教育水平呈负相关关系。

3) 误差选择方面：针对于测试获得的结果，弹性网络效果最好；岭回归次之；Lasso 回归最弱。虽然在训练集上的表现相较于岭回归弱，但将模型应用于训练之外的测试集时，其效果就相对最好。因此弹性网络有较好的外推性。

参考文献

[1] 赵德慧, 胡婷, 谭燕南. 我国教育水平影响因素及提高途径探析[J]. 中国集体经济, 2014(19): 153-154.

[2] Vinod, H.D. (2020) What's the Big Idea? Ridge Regression and Regularisation. *Significance*, **17**, 41 p. https://doi.org/10.1111/1740-9713.01472

[3] 张沥今, 魏夏琰, 陆嘉琦, 潘俊豪. Lasso 回归: 从解释到预测[J]. 心理科学进展, 2020, 28(10): 1777-1791.

[4] Meier, L., Van De Geer, S. and Bühlmann, P. (2008) The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society*: *Series B*, **70**, 53-71. https://doi.org/10.1111/j.1467-9868.2007.00627.x

[5] Tibshirani, R., Bien, J., Friedman, J., *et al*. (2012) Strong Rules for Discarding Predictors in Lasso-Type Problems. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **74**, 245-266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

[6] 谢琍, 唐甜, 王晓瑞. 线性空间自回归模型的不同惩罚函数下参数估计的比较及其实证分析[J]. 数理统计与管理, 2019, 38(5): 823-835.

[7] De Mol, C., De Vito, E. and Rosasco, L. (2009) Elastic-Net Regularization in Learning Theory. *Journal of Complexity*, **25**, 201-230. https://doi.org/10.1016/j.jco.2009.01.002

[8] 曾津, 周建军. 高维数据变量选择方法综述[J]. 数理统计与管理, 2017, 36(4): 678-692.

[9] 王沛立, 李恩平. 我国居民医疗负担及其影响因素分析——基于弹性网方法的实证研究[J]. 数学的实践与认识, 2019, 49(14): 97-107.

[10] 丁澍, 王艳. 高职院校课堂教学质量影响因素研究——基于 Lasso-Logistic 回归模型[J]. 数理统计与管理, 2017, 36(6): 1039-1048.