基于 Logistic 模型在高中文理分科中的应用

尚宏印　冉君丽　杨春松
兴义民族师范学院　数学科学学院，中国·贵州 兴义 562400

【摘　要】普通高中文理课程是一种进一步提高国民素质、面向社会公众的高等基础教育课程，而普通高中文理课程分科是学生未来成长和发展道路上的分岔路，可以充分满足各种潜质的学生自身发展需要。我国高中文理分科并随着新世纪以来高考制度的“3+X”改革而逐渐固化为文理分科教学，使其高中生最后的“准确”选择尤为重要。本文以探讨文理分科的Logistic回归模型为切入点，采用语文、数学，英语等九个定量数据和文科、理科两个定性数据，进行相关性检验、模型系数检验、最大的似然平方对数值的拟合检验等对学生的文理性作出科学的判断和分析。

【关键词】高中文理分科；预测；logistic回归模型
【基金项目】项目名称：大学生职业成熟度对就业现状调查的统计分析，项目编号：20195201868。

1　引言

2010年正式颁布的《国家中长期教育改革和发展规划纲要(2010年-2020年)》[1]对高中文理分科的避而不谈，学界的讨论热潮稍有降温趋势，然而近年来，国家便颁布《国家中长期教育改革和发展规划纲要》[2]向公民收集文理科的存在性意见，可见文理分科不仅是讨论热潮而且对于社会来说还十分重要。本文对我国高中文理分科问题展开分析：以Logistic模型为基础理论，分科为实际情况来进行回归和预测，并提出符合实际且可行的解决方法,有着很切实的意义和价值。我国现在的发展需要的是培养文理兼备的高素质人才，有几个省份已经开始实施“3+X”政策，但对于其余的20多个省份实行这一政策还需要很长的时间，并且我国高中一直实行文理分科改革与发展的路径和策略都是：培养社会所需要的各种人才，即培养知识、能力和个性，注重学生科学和人文素养的融合，实施高考学科分科的政策。

2　基本理论

1840年,数学家Verhulst在malthus模型基础上，加深概念，再进一步研究之后提出了Logistic模型[3]。1925年，英国统计学家尤尔在当年的皇家统计学会的主席致辞中肯定了Verhulst的贡献，也将Verhulst提出的S型曲线称之为Logistic曲线。本文主要运用极大似然估计法对Logistic回归模型的参数进行估计[4]。

假设有n个属于学生的实验样本，观测值记为y_1、y_2、y_3…y_n，设 $p_i = p(y_i = 1|x_i)$ 是一定的条件下 $y_i = 1$ 的条件概率；$1\text{-}p_i = p(y_i = 0|x_i)$ 为 $y_i = 0$ 的条件概率，便得到（2.1）这个观测值的条件概率：

$$p(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}\ ，\ 其中\begin{cases} y_i = 0, 代表学生选文科 \\ y_i = 1, 代表学生选理科 \end{cases} \tag{2.1}$$

对上式两边的取自然对数可以得到（2.3）这个对数似然函数

$$l(\theta) = \ln[L(\theta)] = \ln\prod_{i=1}^{n} p_i^{y_i} \ln\left[\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}\right]$$
$$= \sum_{i=1}^{n}\left\{y_i\left(\beta 0 + \sum_{k=1}^{m}\beta_k x_{ki}\right) - \ln\left[1+\exp\left(\beta_0 + \sum_{k=1}^{m}\beta_k x_{ki}\right)\right]\right\} \tag{2.2}$$

对 $\beta_k(k = 0,1,2,\cdots m)$ 求偏导数，并且让其与0相等，便可得出似然方程：

$$\frac{\partial l(\theta)}{\partial \beta_k} = \sum_{i=1}^{n}\left[y_i - \frac{\exp\left(\beta_0 + \sum_{k=1}^{m}\beta_k x_{ki}\right)}{1+\exp\left(\beta_0 + \sum_{k=1}^{m}\beta_k x_{ki}\right)}\right]，$$

$$x_{ki} = 0, (k = 1,2,\cdots m) \tag{2.3}$$

根据上面的式子可简单地得到 $m+1$ 个似然方程，通过一系列的迭代算法可以计算求解出各个参数的 $\beta_k = 0, (k = 1,2,\cdots m)$ 的释然估计值，由上述式子算出的最终概率，也就是 p_i 的极大似然估计值，他为 $x_{ki} = 0, (k = 1,2,\cdots m)$ 条件下 $y_i = 1$ 或 0 的条件概率的估计。

3　实证分析

3.1 数据的获取和编码

数据来自纳雍一中2017届学生高一期末考试和该届学生高二分班情况，一共有1438名学生，有1名学生缺考（缺失值），运用了分类汇总的方法,将每个在校学生的各科课程成绩、分班情况进行汇总，本文主要采用的9个自变量，包括语文、数学、英语、物理、化学、生物、政治、历史、地理（9个自变量均为离散型变量），因变量文理（分类变量，取值为0或1）均来自于《SPSS统计软件在高中文理分科中的应用》[5]中的指标。其中分科是指文科和理科。

数据的编码按照每一个变量的拼音“首字母”进行编码，语文、数学、英语、物理、化学、生物、政治、历史、地理、分科分别为yw、sx、yy、wl、hx、sw、zz、ls、dl、fk。

3.2 相关性检验

为了进一步研究变量之间的关系，通过Pearson简单相关系数对所有变量进行了检验分析，取显著性水平 α 为0.01时，数学、物理、化学、生物、政治、历史与分科存在相关关系，

拒绝相关系数的原假设 H_0（两变量为零相关），且结果为：-0.364、-0.393、-0.354、-0.227、0.080、0.091，由此可知政治、历史与科目呈正相关。数学、物理、化学、生物与分科呈负相关，数学与分科相关性最高。语文、英语、地理与分科之间不显著，分别为0.006、-0.045、-0.037，所以剔除这三个变量[6]。剔除语文、英语、地理这三个变量后，以分科为因变量做系数检验，得到数学、物理、化学、生物、政治、历史的Sig（P值）分别为0.000、0.000、0.000、0.006、0.000、0.000、0.000，均小于0.05，由此可知，当显著水平 α 为0.01时，应拒绝回归系数检验的原假设，认为数学、物理、化学、生物、政治、历史与分科有显著的线性关系，应保留在此模型中。 3.3模型建立和分析

Logistic建模和其它建模方法有着异曲同工之妙，第一步就是因变量编码，然后寻找最“好”值。以最大似然为基础，使迭代过程逐渐收敛，并达到收敛。当所需的参数达到一定的固定值时，就会得出我们所需要的理想参数值。用初始值经过一系列的检验和分析，就能给出的预测和分类结果。这样的结果主要用于预测前后的结果对比，并朝着良好效果的方向发展。

表3.1　初始方程中的变量表

		B	S.E,	Wald	df	Sig.	Exp(B)
步骤0	常量	-1.455	0.067	466.627	1	0	0.233

从表3.1可以看出该预测的模型是如何给参数赋值的。最初一个开始只是对一个常数项进行赋值，结果为B=-1.455，标准误差的值为S.E.=0.067，Wald值为466.627，自由度（df）=1，sig= 0.00<0.05，模型拟合效果很好，Exp(B)是B还原之后数值，显然 Exp(B)=0.233，因为 x_j 的回归系数 $\beta_j=-1.455<0$，x_j 每次增加1个单位后，结果与增加前的结果相比，事件 $OR_j=0.233<1$，表明 x_j 起到的是保护作用（促进作用）。

表3.2　模型系数的综合检验表

		卡方	df	Sig.
	步骤	6.509	1	0.011
	块	465.239	6	0
	模型	465.239	6	0

表3.2给出了卡方总数值，并且分别给出其数值对应的自由度和P值。给出的显著性水平为0.05，自由度为6，对应的卡方水平临界值分别为12.592。计算出的卡方值大都是低于临界值，而且相应的Sig.值小于0.05，所以，模型系数检验通过。

表3.3　模型汇总表

步骤	-2对数似然值	Cox&SnellR方	NagelkerkeR方
6	929.174[a]	0.277	0.445

表3.3中给出最大似然平方的对数为929.174，主要用途是用于设计实现一个检验模型的对数总体性能和拟数综合检验效果，该对数值在理论上应该是完全服从于似然卡方的对数分布，最大的似然平方值的对数比较大，因此，最大的似然平方对对数值的拟合检验顺利通过。

表3.4　Hosmer和Lemeshow检验

步骤	卡方	df	Sig.
6	10.263	8	0.247

样本数目对似然比函数的自然对数值很重要，作为补充和参照，我们需要进行表3.4中的HL检验[7]。该种卡方检验仍基于传统卡方分布，但是它的检验算法方向和分析结果与目前传统卡方检验的两种方法方式有所很大不同：我们同时需要注意使它的分布卡方值远远低于临界点的值，而非远远高于临界点的值。取显著性水平0.05，由表可知自由度数目df为8，卡方临界值15.50731。作为HL检验的卡方值10.263<15.507，检验通过。Sig值0.247大于0.05，拒接原假设，认为该模型对于拟合数据完全适用。

3.4模型预测

表3.5　最终预测分类表

	已观测		fk	已预测	百分比校正
			理科	文科	
步骤6	fk	理科	1107	58	95.0
		文科	147	125	46.0
	总计百分比				85.7
a.切割值为.500					

表3.5是通过六步迭代进行运算，模型中的参数逐步收敛并达到一个稳定的值，因此我们获取了最终的模型参数。利用最后的 logistic 预测模型，可以进行预测因变量，这便使我们可以清楚地看到，观测值中理科预测有1165个（fk=0），对应预测值有1107个，其预测正确率为95%；而观测值文科有272个（fk=1），对应预测值有125个，其预测正确率为46%。总的预测正确率为85.7%，可以看出模型效果良好。

表3.6　最终方程中的变量

		B	S.E,	Wals	df	Sig.	Exp(B)
步骤6[f]	sx	-0.031	0.006	29.749	1	0	0.969
	wl	-0.064	0.009	54.442	1	0	0.938
	hx	-0.048	0.008	36.896	1	0	0.953
	sw	-0.04	0.016	6.395	1	0.011	0.961
	zz	0.067	0.01	41.265	1	0	1.069
	ls	0.067	0.014	24.084	1	0	1.069
	常量	-3.103	0.794	15.255	1	0	0.045

由表3.6最终得到Logistic模型：

$$p(fk)=\frac{1}{1+e^{-3.103+(-0.031)\times sx+(-0.064\times wl)+(-0.048)\times hx+(-0.040)\times sw+0.067\times zz+0.067\times ls}} \quad (3.1)$$

即：

$$\ln\left(\frac{p(fk)}{1-p(fk)}\right)=-3.103+(-0.031)\times sx+(-0.064\times wl)+(-0.048)\times hx+(-0.040)\times sw+0.067\times zz+0.067\times ls \quad (3.2)$$

将该方程代入数据可以预测每个学生选择文理科的情况，预测值大于0.5说明，学生可能选择文科，预测值小于0.5，说明学生可能选择理科。

4　结论和建议

4.1结论

经过研究分析得到以下结论：

（1）数学与文理分科相关性最高，语文、英语、地理与

文理分科之间不显著，即学生选择文科还是理科与语文、英语、地理成绩高低无关。

（2）观测值理科有1165个，所对应预测值有1107个，得到预测正确率为95%；而观测值文科有272个，所对应预测值有125个,得到预测正确率为46%。最终得到的总的预测正确率为85.7%。物理的B值最低为-0.064，政治和历史的B值最大为0.067，说明物理好的学生偏重于选择理科，而政治和历史好的学生偏重于选文科。

4.2建议

根据以上结论，提出以下建议：

1.高中文理分科的问题使我们认识到，普通的高中课程应该为学生的个人发展提供更多样化的选择。语文、英语、地理与分科选择关系不高，学生在学习该三科课程时，应使出百分之一百二十的努力，让自己该三科的成绩不落后其它同学。

2.单位变量的选取时，应该排除异常的单位变量，异常单位变量的存在会导致R方的值偏小，对可行度有一定的影响。样本量不宜过低，过低会导致预测的正确率不较低。高等教育的平衡，一定要符合跨学科与融合的和谐模式，符合社会产业结构。物理好的同学应注重化学、生物的成绩，政治和历史好的同学应注重生物的成绩。这样可以使现在的高中文理分科为未来的“3+X”模式做好铺垫，为社会培养更多的“高层次”人才。

参考文献:

[1]《国家中长期教育改革和发展规划纲要（2010年-2020年）》[Z].中国法制出版社.

[2]王秀娴.普通高中文理分科的历史问题变迁[EB/OL].文教资料,2016.

[3]姜爱平,郝慧娟.Logistic模型参数估计研究[J].海南师范大学学报,海南:,2014.27(4).

[4]朱华锋.Logistic模型的参数估计及其实证研究分析[J].高校讲坛:2011.(1):169-170.

[5]谭志昌.SPSS统计软件在高中文理分科中的应用[J].山西师范大学学报(自然科学版)研究生论文专刊,2009,23(2):1-3.

[6]闫成海.校园超市购物满意度logistic模型研究——以某高校为例[J].数学实践与认识,2019,49(15):3-4.

[7]陈彦光.研究生地理数学方法(实习)[M].北京:北京大学环境学院,2006.182-190.