

基于因子分析的高中生成绩评价研究

弓瑞洁

长安大学理学院, 陕西 西安

Email: 1572074842@qq.com

收稿日期: 2020年9月29日; 录用日期: 2020年10月14日; 发布日期: 2020年10月21日

摘要

高中学生成绩作为教学管理的重要内容, 是评价高中生综合素质的重要数字化指标。它可以反映学生掌握知识、各方面能力以及教师教学水平的程度, 也为学生未来的发展提供了一项重要参考因素。因此, 对学生成绩综合评价的数学模型研究有益于对学生成绩做出准确全面的评价。因子分析综合评价模型是在因子分析基础下对学生成绩进行综合评价的一种模型, 通过对学生成绩进行主成分分析, 借助于SPSS软件应用因子分析法, 从而计算学生的因子综合得分并进行得分排序。这种综合评价模型可以弥补一些高中直接计算学生总成绩进行排名的缺点, 评价结果更为合理、公平、科学。

关键词

因子分析, 降维, 综合评价

Research on High School Student Achievement Evaluation Based on the Factor Analysis

Ruijie Gong

Department of Science, Chang'an University, Xi'an Shaanxi

Email: 1572074842@qq.com

Received: Sep. 29th, 2020; accepted: Oct. 14th, 2020; published: Oct. 21st, 2020

Abstract

High school student grades, as an important content of teaching management, are an important digital index to evaluate the comprehensive quality of high school students. It can reflect students'

文章引用: 弓瑞洁. 基于因子分析的高中生成绩评价研究[J]. 统计学与应用, 2020, 9(5): 784-791.

DOI: 10.12677/sa.2020.95081

mastery of knowledge, various abilities and teachers' teaching level, and also provides an important reference factor for students' future development. Therefore, the mathematical model research on the comprehensive evaluation of students' performance is beneficial to make accurate and comprehensive evaluation of students' performance. Factor analysis comprehensive evaluation model is a kind of comprehensive evaluation model for students' scores based on factor analysis. Through the principal component analysis of students' scores and the application of factor analysis with the help of SPSS software, the comprehensive factor scores of students are calculated and the scores are ranked. This comprehensive evaluation model can make up for the shortcomings of some high school students' ranking by directly calculating their total scores, and the evaluation results are more reasonable, fair and scientific.

Keywords

Factor Analysis, Dimensionality Reduction, Comprehensive Evaluation

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

因子分析是最早是用来解决智力测验得分的,对于解决智力测验起到重要作用。因子分析是利用降维思想来简化数据的一种多元统计方法。它的本质可以这样理解:它从原始变量的相关矩阵出发,通过实际数据,进行分析后,利用几个“抽象”变量来表示其基本的数据结构。这几个“抽象”变量称为“因子”,因子可以代替原始众多变量,用公因子反映原始众多变量的大部分信息[1]。本文根据河北省某高中一年级学生的各科成绩的实际数据,经过大量的统计后,进行因子分析得到学生的成绩综合评价模型,在此基础上将其与常用的总成绩排名法进行比较分析与评价,结果一致认为该方法明显有其独特的优势,它的优势可弥补平均成绩法的不足。通过对比,我们可以发现学生的个体特征和群体特征,可以看出它具有较强的科学性,是一种衡量学生成绩行之有效的方法。

2. 因子分析的基本理论

因子分析是用少数几个特殊因子去研究多个原始指标间相关关系的一种多元统计方法。统计方法比一般方法占优,我们从根本出发看下它的定义,因子分析可分为R型因子分析和Q型因子分析,两者计算过程是一样的,但出发点不同,前者是从变量间的相关系数矩阵出发来研究变量之间的相关关系,后者是从样品间的相似系数矩阵出发来研究样品间的相关关系。后者更能说明因子分析的基本理论[2][3]。R型因子分析通常将各个变量表示成公因子的线性函数与特殊因子之和(公式1),也可以用矩阵形式表示为: $X = AF + \varepsilon$, F 称为公共因子, ε 称为 X 的特殊因子。 a_{ij} 称为第 i 个变量在第 j 个因子上的负荷,也表示 m 维空间中一点 X_i 在坐标轴 F_j 上的投影, A 称为因子载荷矩阵。

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{ij}F_j + \cdots + a_{im}F_m + \varepsilon_i, \quad i = 1, 2, \cdots, p; \quad j = 1, 2, \cdots, m; \quad (1)$$

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pm} \end{bmatrix} = (A_1, A_2, \cdots, A_m), \quad m \leq p. \quad (2)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \quad (3)$$

在正交因子模型的假定下, 随机向量 X 的协方差阵 Σ 要分解成两部分, 但这种分解并不是唯一的。设 T 为一个 $m \times m$ 正交矩阵, 则 $TT' = T'T = I$ 。于是,

$$\Sigma = ATT'F + \Phi \quad (4)$$

若令 $A^* = AT$, $F^* = T'F$, 则模型可表示为: $X = A^*F^* + \varepsilon$, 且满足因子模型的条件:

- (1) $E(F^*) = E(T'F) = T'E(F) = 0$, $\text{Var}(F^*) = \text{Var}(T'F) = T'\text{Var}(F)T = I_m$;
- (2) $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \text{diag}(\Phi_1, \Phi_2, \dots, \Phi_p)$;
- (3) $\text{cov}(F^*, \varepsilon) = \text{cov}(T'F, \varepsilon) = T'\text{cov}(F, \varepsilon) = 0$ 。

公因子 F 不是唯一的, 因子载荷矩阵 A 也不是唯一的。它们是可以随机的变化的, 只要对公因子作正交变换, 就可以得到新的公因子。在几何上, 正交变换对应坐标轴的旋转。旋转是在空间上旋转的, 坐标系旋转后, 因子载荷也发生相应变化, 因此因子载荷矩阵不唯一, 具有多样性。

$$\text{cov}(x_i, f_j) = \text{cov}(a_{i1}f_1 + a_{i2}f_2 + \dots + a_{im}f_m + \varepsilon_i, f_j) = \sum_{k=1}^m a_{ik} \text{cov}(f_k, f_j) \quad (5)$$

当公因子之间完全不相关时, 即 $\text{cov}(f_k, f_j) = 0 (k \neq j)$, 而 $\text{cov}(f_j, f_j) = \text{Var}(f_j) = 1$, 此时, $\text{cov}(x_i, f_j) = a_{ij} = r_{x_i, f_j}$ 。当公因子之间完全不相关时, a_{ij} 就是第 i 个变量和第 j 个公因子之间的相关系数, 即 x_i 在第 j 个公因子上的相对重要性。 a_{ij} 的绝对值越大, 表示公因子 f_j 与变量 x_i 的关系越密切, 可以据此去寻找公因子 f_j 的实际含义。

$$r_{x_i, f_j} = \text{cov}(x_i, x_j) = \text{cov}\left(\sum_{k=1}^m a_{ik}f_k, \sum_{k=1}^m a_{jk}f_k\right) = \sum_{k=1}^m a_{ik}a_{jk} \quad (6)$$

各变量的共同度是衡量因子分析效果的一个重要指标。变量共同度反映的是 m 个公因子对原始变量 x_i 的总方差解释的比例。设因子载荷矩阵为 A , 变量 X_i 的共同度是第 i 行元素的平方和

$$h_i^2 = \sum_{j=1}^m a_{ij}^2 \quad (7)$$

$$\text{Var}(x_i) = \sum_{j=1}^m \text{Var}(a_{ij}f_j) + \text{Var}(\varepsilon_i) = \sum_{j=1}^m a_{ij}^2 + \sigma_i^2 = h_i^2 + \sigma_i^2 \quad (8)$$

上式说明变量 x_i 的方差由两部分组成: 第一部分为共同度 h_i^2 , 它描述了全部公共因子对变量 x_i 的总方差所作的贡献, 反映了公共因子对变量 x_i 的影响程度。第二部分为特殊因子 ε_i 对变量 x_i 的方差的贡献, 通常称为个性方差, 如果对 x_i 作了标准化处理, 有

$$1 = h_i^2 + \sigma_i^2 \quad (9)$$

公因子 f_i 的方差贡献, 等于和该因子有关的因子载荷的平方和, 即

$$g_j^2 = \sum_{i=1}^p a_{ij}^2 \quad (10)$$

公因子方差贡献所做的贡献, 反映了该因子对所有原始变量总方差的解释能力, 是衡量公因子相对重要性的指标参数, 指标参数越高, 越能说明该值越高说明公因子的重要程度越高。在因子分析模型 $X = AF + \varepsilon$ 中, 如果不考虑特殊因子的影响, 当 $m = p$ 且 A 可逆时, 我们可以非常方便的从每个样品的

指标取值 X 计算出其在因子 F 上的相应取值: $F = A^{-1}X$, 即该样品在因子 F 上的“得分”情况, 也可以简称为该样品的因子得分。

但是因子分析模型在实际应用中要求 $m < p$, 因此, 不能精确计算出因子的得分情况, 只能对因子得分进行估计, 计算它的一个估计值, 汤姆森回归法假设公共因子可在对 p 个原始变量作回归,

$$\hat{F}_j = b_{j0} + b_{j1}X_1 + \cdots + b_{jp}X_p, \quad (j=1, \cdots, m) \quad (11)$$

如果 F_j, x_i 都标准化了, 回归的常数项为零, 即 $b_{j0} = 0$ 。由因子载荷的统计意义知道, 对于任意的 $i = 1, \cdots, p; j = 1, \cdots, m$ 都有:

$$\begin{aligned} a_{ij} = r_{X_i, F_j} &= E(X_i F_j) = E[X_i (b_{j1}X_1 + \cdots + b_{jp}X_p)] \\ &= b_{j1}E(X_i X_1) + \cdots + b_{jp}E(X_i X_p) = b_{j1}r_{i1} + \cdots + b_{jp}r_{ip} \end{aligned} \quad (12)$$

$$\text{记 } B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix}, \text{ 矩阵形式可表示为 } A = RB' \text{ 或 } B = A'R^{-1}$$

$$\hat{F} = \begin{bmatrix} \hat{F}_1 \\ \vdots \\ \hat{F}_m \end{bmatrix} = \begin{bmatrix} b'_1 X \\ \vdots \\ b'_m X \end{bmatrix} = BX = A'R^{-1}X \quad (13)$$

可以得到因子得分的估算公式为: $\hat{F} = A'R^{-1}X$, 其中 R 是 X 的相关系数矩阵。

3. 因子分析的一般步骤

1) 要想用因子分析法解决一个问题, 首先要利用 KMO 检验和 Bartlett 球形检验对数据进行检验。如果 KMO 值越接近于 1, 越适合做因子分析。如果 KMO 值小于 0.5, 则不适合做因子分析。当 Bartlett 检验统计量 p 值小于 0.05 时, 则变量适合做因子分析。

2) 一般利用主成分分析的方法是提取公因子。

3) 对样本进行因子分析, 通过分析因子的方差贡献率信息和旋转后的因子载荷分布, 来确定变量和因子间相关关系。

4) 计算因子综合得分, 我们可以直观量化结果来确定样本的综合排序。

4. 实证研究

高中学生综合成绩对评价教学质量有很大作用, 检验课程设置和人才培养合理性等具有非常重要的参考价值。目前高中生成绩评定一般是直接按照各科成绩总分进行排名。这种方法简单易行, 直观易于理解, 动手操作简单, 但太过片面笼统, 有明显的缺点, 不能反映学生的特长与个性差异, 不能看出学生的实践能力, 不能全面评价学生的综合素质揭示群体特征, 因此这种评价方法有一定弊端[4][5]。本文为评价学生成绩提供了另一种方法——因子分析法。这种方法弥补了以上缺点, 从数据的内部结构出发, 理论联系实际, 挖掘影响学生成绩的潜在因子, 其能更加客观、科学地解释影响学生成绩的因素, 从而更好地指导于日常教学与学生培养。本文建立了学生一个更为科学的综合成绩评价模型, 从而挖掘影响学生成绩的内在规律。

本文以河北省某高中 100 名高一年级学生的各科成绩为原始数据, 选取了 13 门课程, 分别是: 语文(X_1)、数学(X_2)、英语(X_3)、物理(X_4)、化学(X_5)、生物(X_6)、政治(X_7)、历史(X_8)、地理(X_9)、音乐(X_{10})、美

术(X_{11})、体育(X_{12})、信息技术(X_{13})。借助 SPSS 对原始数据进行因子分析,找出隐含的潜在因子来解释学生的成绩进而科学合理地指导日常教学并提高学生的综合素质。

(一) 适宜性检验。本文利用 SPSS 对处理后的数据进行 KMO 检验和巴特利特球形(如表 1), 检验结果为 $KMO = 0.578 > 0.5$, 勉强适合做因子分析。 $P = 0.000 < 0.05$ 这表明样本取样度合理, 变量间的相关性较强, 适合做因子分析。

Table 1. KMO and Bartlett's test

表 1. KMO 和 Bartlett 的检验

KMO 检验	巴特利特球形度检验		
	近似卡方	自由度	显著性
0.578	198.260	78	0.000

(二) 公因子选取与解释。本文用主成分分析法, 选取特征值大于 1 的 5 个公因子。表 2 是各个公因子对于总方差的解释程度, 其累计方差贡献率为 61.592%, 能够反映原始数据的大部分信息。即利用因子分析的方法将原问题中 13 门科目指标变量通过 5 个公共因子代替, 对样本数据做到了较大程度的降维。

Table 2. Total variance explained

表 2. 特征根与方差贡献率表

因子	初始特征值			旋转后载荷平方和		
	特征值	方差贡献率(%)	累积贡献率(%)	特征值	方差贡献率(%)	累积贡献率(%)
1	2.560	19.692	19.692	2.309	17.758	17.758
2	1.686	12.968	32.660	1.571	12.083	29.841
3	1.444	11.110	43.770	1.430	11.001	40.842
4	1.252	9.631	53.401	1.393	10.714	51.557
5	1.065	8.192	61.592	1.305	10.036	61.592
6	0.915	7.037	68.629			
7	0.895	6.883	75.512			
8	0.688	5.289	80.801			
9	0.668	5.140	85.941			
10	0.597	4.595	90.536			
11	0.531	4.084	94.620			
12	0.368	2.828	97.448			
13	0.332	2.552	100.000			

因子载荷则表示公因子与原始变量之间的相关性。初始因子载荷解释不够明确, 通过正交旋转法计算出旋转后因子载荷矩阵(如表 3)。旋转后可以看到, 第一个因子在变量 X_1, X_2, X_3, X_4 上有较大载荷, 可

将第一个因子命名为综合能力因子。第二个因子在变量 X_{10}, X_{11} 上有较大载荷, 可将第二个因子命名为艺术水平因子。第三个因子在 X_5, X_6 上有较大载荷, 可将第三个因子命名为理科能力因子。第四个因子在 X_8, X_9, X_{12} 上有较大载荷, 可将第四个因子命名为文科及体育因子。第五个因子在 X_7, X_{13} 上有较大载荷, 可将第五个因子命名为哲学及计算机能力因子。

Table 3. Rotated factor matrix
表 3. 旋转后的因子载荷阵

	因子				
	1	2	3	4	5
语文(X_1)	-0.678	0.225	-0.146	-0.184	0.036
数学(X_2)	0.705	-0.081	0.166	-0.232	-0.156
英语(X_3)	-0.602	-0.152	-0.155	-0.323	0.176
物理(X_4)	0.851	-0.011	-0.248	-0.043	0.042
化学(X_5)	0.293	0.143	0.594	0.053	0.166
生物(X_6)	-0.061	-0.141	0.818	0.056	-0.108
政治(X_7)	-0.217	0.254	0.319	-0.072	0.686
历史(X_8)	-0.093	0.145	0.075	0.800	-0.021
地理(X_9)	0.307	-0.103	0.249	0.416	-0.058
音乐(X_{10})	0.069	-0.747	0.102	-0.193	0.195
美术(X_{11})	-0.048	0.775	0.040	-0.135	0.113
体育(X_{12})	0.121	-0.378	-0.256	0.561	0.311
信息技术(X_{13})	0.057	0.227	0.163	-0.071	-0.765

表 4 为旋转后的因子载荷矩阵, 该表通过因子得分表达式计算样本在单一因子上的表现效果。以因子得分结果作为替代指标近似值, 以因子方差贡献率和累计方差贡献率的比值作为替代指标的权重, 可以计算得到每个样本的综合得分作为评价学生综合成绩的依据。

$$F_1 = -0.272X_1 + 0.317X_2 - 0.244X_3 + 0.432X_4 + 0.131X_5 - 0.117X_6 - 0.007X_7 - 0.103X_8 + 0.078X_9 - 0.009X_{10} + 0.078X_{11} + 0.031X_{12} - 0.052X_{13}$$

$$\begin{aligned} F_2 &= 0.089X_1 - 0.015X_2 - 0.147X_3 + 0.097X_4 + 0.105X_5 - 0.161X_6 + 0.189X_7 \\ &\quad + 0.107X_8 - 0.046X_9 - 0.489X_{10} + 0.519X_{11} - 0.178X_{12} + 0.074X_{13} \\ F_3 &= -0.066X_1 + 0.085X_2 - 0.049X_3 - 0.233X_4 + 0.394X_5 + 0.597X_6 + 0.229X_7 \\ &\quad + 0.022X_8 + 0.148X_9 + 0.122X_{10} - 0.012X_{11} - 0.189X_{12} + 0.102X_{13} \\ F_4 &= -0.069X_1 - 0.230X_2 - 0.198X_3 - 0.095X_4 - 0.010X_5 + 0.017X_6 - 0.074X_7 \\ &\quad + 0.603X_8 + 0.273X_9 - 0.187X_{10} - 0.078X_{11} + 0.386X_{12} - 0.017X_{13} \\ F_5 &= -0.027X_1 - 0.032X_2 + 0.069X_3 + 0.144X_4 + 0.181X_5 - 0.112X_6 + 0.554X_7 \\ &\quad - 0.056X_8 - 0.039X_9 + 0.107X_{10} + 0.164X_{11} + 0.204X_{12} - 0.588X_{13} \end{aligned}$$

Table 4. Factor score coefficient matrix
表 4. 因子得分系数矩阵

	因子				
	1	2	3	4	5
语文(X_1)	-0.272	0.089	-0.066	-0.069	-0.027
数学(X_2)	0.317	-0.015	0.085	-0.230	-0.032
英语(X_3)	-0.244	-0.147	-0.049	-0.198	0.069
物理(X_4)	0.432	0.097	-0.233	-0.095	0.144
化学(X_5)	0.131	0.105	0.394	-0.010	0.181
生物(X_6)	-0.117	-0.161	0.597	0.017	-0.112
政治(X_7)	-0.007	0.189	0.229	-0.074	0.554
历史(X_8)	-0.103	0.107	0.022	0.603	-0.056
地理(X_9)	0.078	-0.046	0.148	0.273	-0.039
音乐(X_{10})	-0.009	-0.489	0.122	-0.187	0.107
美术(X_{11})	0.078	0.519	-0.012	-0.078	0.164
体育(X_{12})	0.031	-0.178	-0.189	0.386	0.204
信息技术(X_{13})	-0.052	0.074	0.102	-0.017	-0.588

(三) 综合评价。通过因子得分系数矩阵, 计算出因子得分函数, 并得到学生在各公因子中的得分, 在此基础上以各公因子的方差贡献率为权重并利用线性组合建立学生成绩综合模型, 其模型如下:

$$S = 0.1776F_1 + 0.1208F_2 + 0.11001F_3 + 0.10714F_4 + 0.10036F_5$$

Table 5. Comprehensive evaluation
表 5. 综合评价

学号	F_1	F_2	F_3	F_4	F_5	因子综合得分	因子综合得分排名	总成绩排名
1	43.039	26.709	96.3	20.244	33.532	26.998	1	1
26	41.65	20.191	91.521	27.248	37.678	26.605	2	27
20	27.056	35.395	95.841	17.588	46.462	26.172	3	24
9	43.167	21.126	85.709	21.369	41.138	26.065	4	15
2	39.027	4.607	91.709	30.786	51.362	26.029	5	4
5	56.927	17.983	89.937	2.179	35.577	25.989	6	5
16	35.73	24.915	98.862	23.335	30.131	25.755	7	14
18	54.538	16.069	79.739	18.385	31.251	25.505	8	11
12	37.717	29.815	80.001	22.956	38.956	25.470	9	17
21	42.624	28.397	97.394	1.368	34.738	25.348	10	25

从表 5 的综合评价中可以看出, 通过因子分析法和直接计算学生总成绩这两种方法比较后前者具有明显优势, 如: 26 号学生因子得分综合排名为第 1 名, 但是总成绩排名则为第 27 名。详细观察发现该生在第四主成分上存在明显的优势, 可见该生文科能力很强。第一主成分和第三主成分上也优势明显,

因此综合能力和理科能力也突出。21号学生其因子得分排名为第10名，但是总成绩排名则为第25名。通过分析发现该生在第三主成分上有明显的优势，在第四主成分上存在明显的不足，即文科能力水平明显偏低，但是理科能力表现优异，所以综合得分排名第10名，而平均成绩排名却相当靠后。可见因子分析有其独特的优势，可发现学生的个性特征和综合能力，从而更好地因材施教、提高教学质量。这是总成绩排名所不能比拟的，评价结果更为客观，对高中生文理选科也起到了一定的指导作用，同时在新高考改革中，也为学生在选择高考科目时提供了一定的参考作用[6]。

5. 结论与展望

本文在实际的撰写中，利用现有的学生成绩数据同时结合教学的实际情况、学生的不同背景将因子分析的方法应用于实际数据分析中，从多种课程间提取出有价值的信息，如课程之间的相关性、课程之间的归类合并、提取影响学生成绩的潜在因子，发现学生课程学习的差异等，这对学校课程设置、安排以及实际的教学与管理、学生发展等具有非常重要的意义，同时也为教育教学改革提供现实理论基础。

本文基于因子分析的高中学生成绩数据，从统计学的角度，为综合评价学生成绩提供了一种简便的方法。基于河北省高一学生成绩的分析结果可以看到，变量由13降维到5，很好地实现了数据降维的效果，并且得到高中学生综合成绩排名，与常用的总成绩排名方法进行了对比，发现了因子分析综合评价优于常见的学生成绩方法。然而因子分析仍存在问题，例如，对于旋转后因子的命名未能给出统一的解释。在进行成绩综合评价时，因子及其显著相关的原始变量有内在的相关关系，综合因子是原始变量的线性组合，用这些关系对综合因子、逐个因子的变量组进行深入的数据分析，尽可能深入到决策的相关性程度。

参考文献

- [1] 魏劲如. 基于因子分析的基金平价研究[J]. 统计学与应用, 2019, 81(1): 31-38.
- [2] 汪冬华. 多元统计分析与 SPSS 应用[M]. 上海: 华东理工大学出版社, 2010: 208-218.
- [3] 林海明. 因子分析应用中一些常见问题的解析[J]. 统计与决策, 2012(15): 65-69.
- [4] 王小丽, 李林芝, 简太敏. 多元统计分析在大学生综合成绩评价中的应用[J]. 产业与科技论坛, 2018, 17(12): 117-119.
- [5] 刘访华, 余瑞君. 基于因子分析的学生成绩评价对提高本科教学质量的启示[J]. 中国人民大学教育学报, 2013, 11(4): 15-21.
- [6] 张启贤, 陈欣, 刘新平. 基于因子分析下的学生成绩综合评价模型研究[J]. 西安文理学院学报(自然科学版), 2008(11): 1-6.