21-22-2 数学建模与数学实验
课程论文二

题目： 基于理想解法的院校评价模型

组队编号： 2201017

学生 1

姓名 山有虎 **班级** 统计 2 班 **学号** 1111111111

学生 2

姓名 虎山行 **班级** 统计 1 班 **学号** 2222222222

stulink 大学

2022 年 4 月

基于理想解法的院校评价模型

摘　要

本文首先对题目所给数据进行了变量的相关性检查与分析，并根据客观经验对数据进行了异常值处理；其次，本文根据相关资料中世界一流大学的生师比分布情况拟合出了其近似的偏态分布，然后由其分布对数据中生师比属性的值进行了对应的转换以使最终评价更为客观。

模型建立方面，本文采用了两种标准化方法与三种客观赋权法，相应总共建立了五个模型（$2 \times 3 = 6$，本应是六个模型，但经检验，其中一个模型所采用的标准化方法与所采用的加权法不相容，故将其去除），紧接以理想解法分别求出了其相应院校排名，然后计算出各个院校在五次排名结果中的平均得分，最后以这个平均得分的排名作为本文的最终评估结果。

关键词: 相关分析　分布拟合　标准化　客观赋权法　理想解法

一、问题重述

1.1 问题背景

为客观评价我国研究生教育实际状况与各研究生院教学质量，需要先抽取小样本进行试评估。所收集的数据如表 1

1.2 问题提出

试对样本中 5 个院校实力进行评估，最终给出院校排名。

表 1: 研究生院的试评估数据资料

人均专著（本/人）	生师比	科研经费（万元/年）	逾期毕业率
0.1	5	5000	4.7
0.2	6	6000	5.6
0.4	7	7000	6.7
0.9	10	10000	2.3
1.2	2	400	1.8

二、问题分析

本题为典型的小样本评估问题，对其分析如下：

1. 所给样本数据的四个属性中，人均专著与科研经费带有不同的量纲，应在模型的建立中进行规范化（去量纲）处理；

2. 为之后方便比较，需要将各属性数据进行不同的加权操作，而具体权数有待后续确定；

3. 就类型而言，人均专著与科研经费为效益型属性，逾期毕业率为成本型属性，而生师比的类型也待后续进行确定。

三、模型假设

1. 所给数据具有一定的代表性；

2. 本次评估的标杆为世界一流院校。

四、符号说明

符号	说明
a_{ij}	原数据表中第 i 行第 j 列元素的值
b_{ij}	标准化后数据表中第 i 行第 j 列元素的值
w_j	标准化后数据表中第 j 列元素的权重
c_{ij}	加权后的数据表中第 i 行第 j 列元素的值
σ_i	当前数据表第 j 列的样本标准差

五、模型的建立与求解

5.1 结构与思路

本节结构与相关思路如图 1 所示。

5.2 相关分析

本文首先对数据中各属性指标间的相关性进行了可视化，如图 2 。

图 2 显示，院校生师比与科研经费的相关系数高达 0.991。然后经计算得出各院校人均的科研经费依次（从第一行数据所代表院校到最后一行数

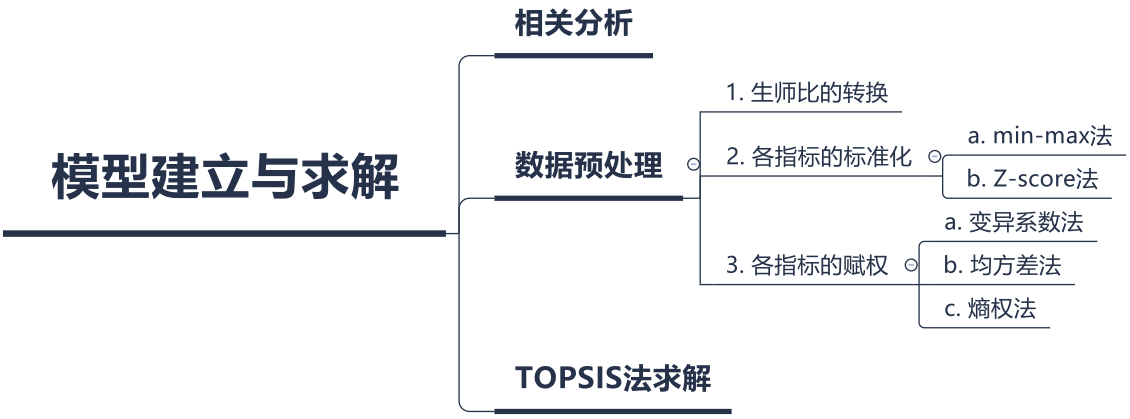

图 1: 模型建立与求解的逻辑图

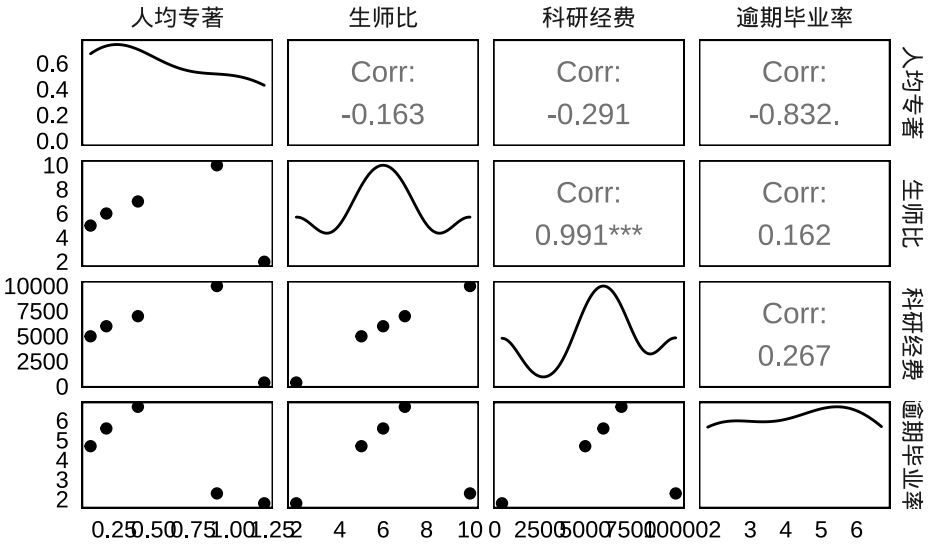

图 2: 原数据各属性指标间的相关性

表 3: 剔除第五所院校后的数据

人均专著	生师比	科研经费	逾期毕业率
0.1	5	5000	4.7
0.2	6	6000	5.6
0.4	7	7000	6.7
0.9	10	10000	2.3

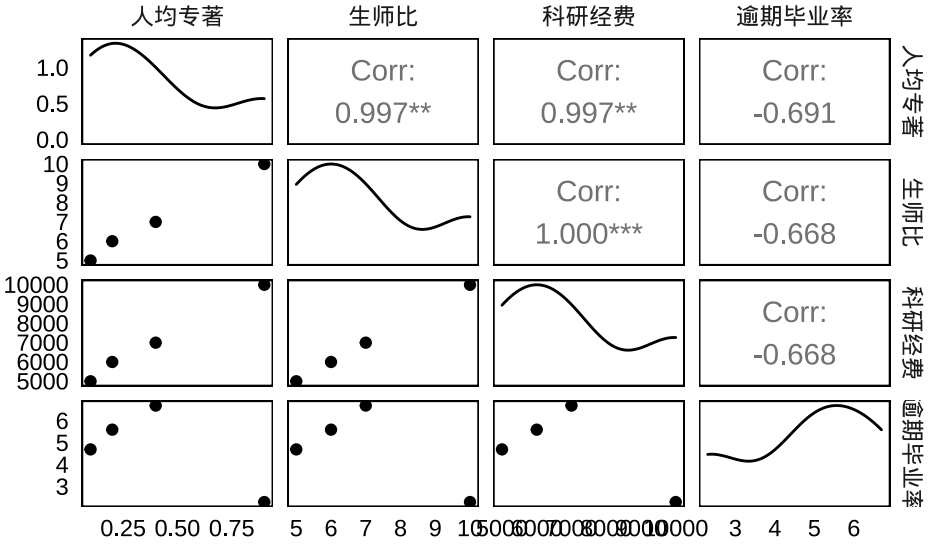

图 3: 剔除第五行数据后各属性间的相关性

据所代表院校）为 1000、1000、1000、1000、200 万元/年。由此可知第五行数据相较其他更为特殊。观察原数据表能够发现第五个院校除了科研经费较少外，其余各项属性指标均非常优异，大大强于另外四者。可以猜测此学校或为文科类院校。由国内大部分院校的首要任务为教书育人，而非制造众多科研项目，故本文首先采取主观判断法，预先将第五个院校定为院校排名的第一名，并在后续的分析数据中将其剔除，只在最后下结论时将其添进排名。

剔除第五个院校的数据后留存的数据 (单位已略去) 如表 3 .

再次对该数据的相关性进行可视化，结果如图 3 .

可以看出在剔除第五行数据后，生师比与人均专著展现出了极高的相关性。为探究其相关关系，本文对这两个属性指标进行了线性回归拟合，可视化结果如图 4 .

由图像可以看出两属性的线性回归拟合得很好。正由其拟合较好，倘若这些样本又具有代表性，那么以后便可仅根据一个学校生师比的值推测

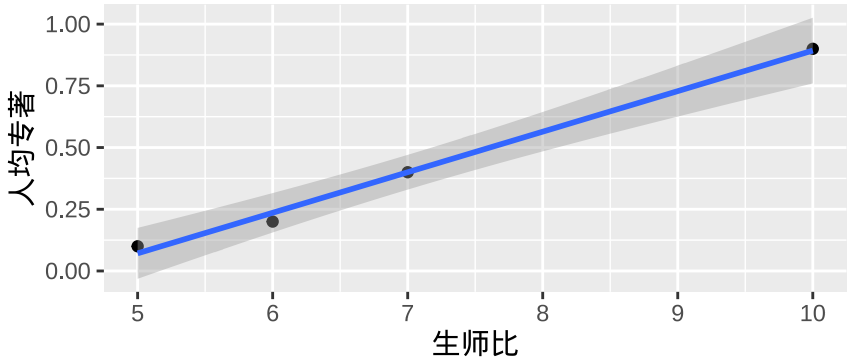

图 4: 人均专著与生师比的回归拟合

出此学校人均专著和科研经费的值。更多相关讨论放在了模型的推广与改进一节。

5.3 数据预处理

在所给数据当中，人均专著和科研经费显然是效益型属性，即相同条件下，它们的值越大，代表该院校实力越强，而逾期毕业率则显然是成本型属性，在院校的评估中，其值越小越好。而生师比这一属性指标无法简单直接地看出它对院校评估所应起到的影响，本节的第一部分就包含了对其的探讨与相应值的转换。

数据中人均专著与科研经费拥有迥然不同的量纲，若要公平地进行后续评估，预先对它们进行标准化处理是很有必要的，本节的第二部分就根据相关资料 [2] 给出了两种标准化的方法，预备后续建模采用。

在进行标准化过后，应当考虑的问题就是各项指标的权重，即要确定人均专著、生师比、科研经费和逾期毕业率这四项指标中，哪个对院校评估的影响最大，哪个指标是次级重要的，以此类推。所要做的具体工作就是以数值化的各指标权重将其对应值做相应转换，同样的，本节的第三部分根据相关资料 [2][3] 给出了三种客观赋权的方法，以备后续求解之用。

5.3.1 生师比的转换

查阅相关资料 [1]，国际一流大学的生师比所在区间一般为 2 到 4，就目前所剩数据（剔除第五行后）而言，无一达到。而该资料 [1] 也表明，生师比在 10 以内都可以接受。资料 [1] 中有世界一流院校的生师比数据（七十四个）统计情况如图 5 .

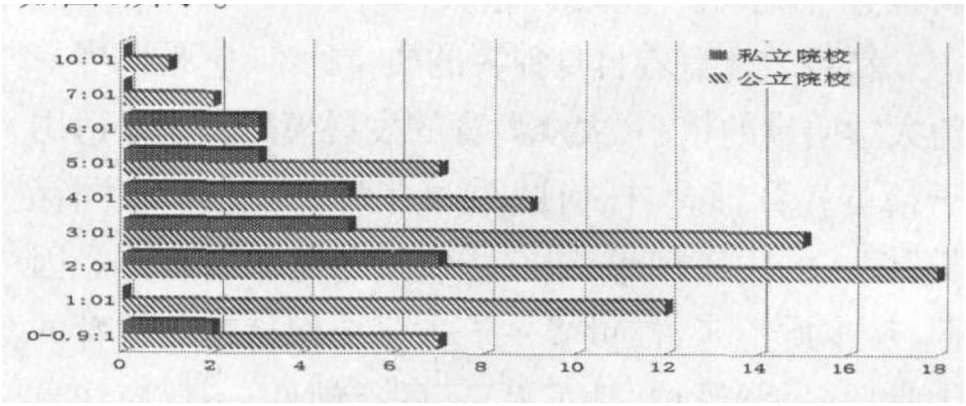

图 5: 一流大学研究生生师比分类统计图示

表 4: 转换生师比后的数据

人均专著	生师比	科研经费	逾期毕业率
0.1	0.074	5000	4.7
0.2	0.042	6000	5.6
0.4	0.023	7000	6.7
0.9	0.003	10000	2.3

可以看出（本文只参考了统计图中公立院校的数据），一流大学的生师比集中在 2 左右，为偏态分布。为确定本文模型生师比加权系数，本文根据参考资料对世界一流大学的调查统计图进行了数据复现并对其进行了伽马分布拟合，拟合结果其基本服从 $X \sim Ga(2.256, 0.797)$，经检验，其 P 值小于 0.01，通过显著性检验。拟合效果如图 6 .

图 6 中，纵坐标 $p(x)$ 为 $X \sim Ga(2.256, 0.797)$ 的密度函数，$p(x)$ 越小，则生师比 x 的值所对应的学校为一流学校的概率越小。基于此，本文将数据中四个院校生师比值所对应的 $p(x)$ 的值赋予数据的生师比列，以达到为其不同值赋予不同权重的目的。显而易见，被赋予 $p(x)$ 后的生师比成为了效益型属性，其将在后面模型建立中参与影响求解结果。

转换生师比后的数据如表 4 .

5.3.2 各指标的标准化

由于本文所采用标准化方法与加权方法数量较多，故本文首先给出所用到的数据处理方法（两种标准化方法与三种客观赋权法）的公式与简单说明，然后将仅在模型一中对模型的建立与求解进行详细描述，而对于其余模型的建立与求解，后续只需带入相应公式即可得出结果，为减少不必

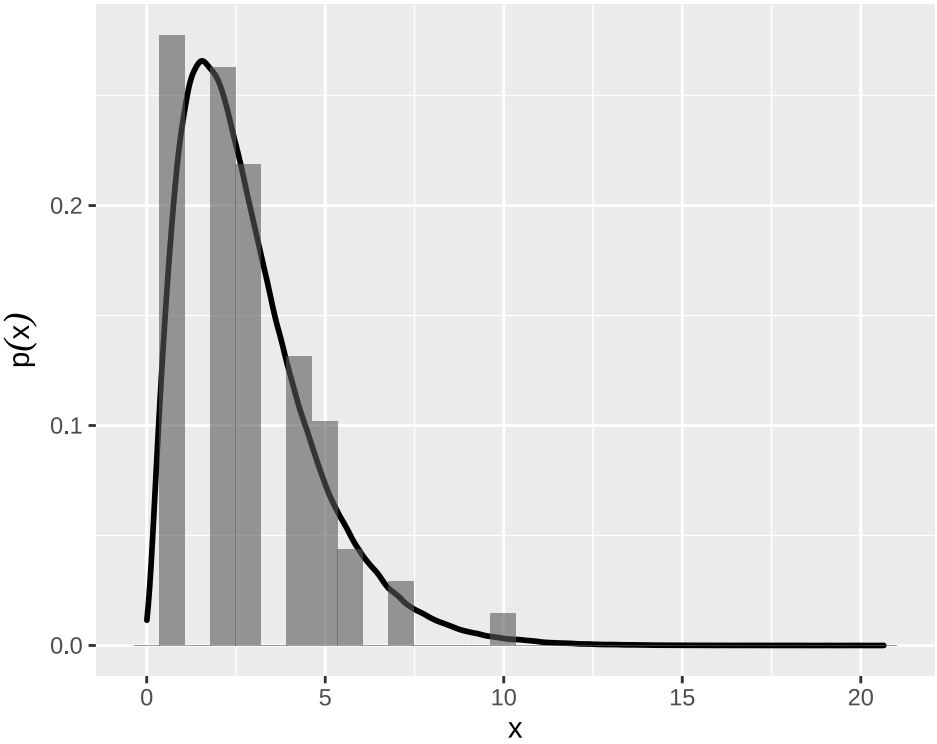

图 6: 一流大学生师比的伽马分布拟合

要的篇幅，本文将略去后续那些繁复的叙述，只给出最终计算结果供读者比较。

本文所采用的两种标准化方法与对应公式分别为（注：若不做说明，$i = 1, \cdots, n \quad j = 1, \cdots, m$ 对以下所有公式皆适用）：

a. min-max 标准化法：

$$b_{ij} = \frac{a_{ij} - \min a_j}{\max a_j - \min a_j}. \tag{1}$$

b. Z-score 标准化法：

$$b_{ij} = \frac{a_{ij} - \overline{a}_j}{\sigma_j} \tag{2}$$

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - \overline{a}_j)^2}. \tag{3}$$

5.3.3 各指标的赋权

本文所采用的三种客观赋权法与对应公式分别为：

a. 变异系数法

$$w_j = \frac{v_j}{\sum_{j=1}^{m} v_j} \tag{4}$$

$$v_j = \frac{\sigma_j}{\overline{b}_j} \tag{5}$$

v_j 为第 j 列数据的变异系数。

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(b_{ij} - \overline{b}_j)^2}. \tag{6}$$

b. 均方差法：

$$w_j = \frac{\sigma_j}{\sum_{j=1}^{m} \sigma_j} \tag{7}$$

同公式 (5)：

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(b_{ij} - \overline{b}_j)^2}.$$

c. 熵权法：

$$p_{ij} = \frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}}. \tag{8}$$

p_{ij} 为第 j 项指标下第 i 个样本值占该指标的比重。

$$e_j = -k\sum_{i=1}^{n} p_{ij} \ln p_{ij}. \tag{9}$$

e_j 为第 j 项指标的熵值。其中，$k = \frac{1}{\ln n > 0}$.

$$d_j = 1 - e_j \tag{10}$$

d_j 为信息熵冗余度（差异）。

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \tag{11}$$

赋权公式

$$c_{ij} = w_j \cdot b_{ij}. \tag{12}$$

5.4 TOPSIS 法求解

TOPSIS 法也称为理想解法，是一种有效的多指标评价方法。

它首先需要构造评价问题的正理想解和负理想解（即各个指标的最优，最劣解），然后通过计算每个指标对正理想解的靠近程度，对负理想解的远离程度来对院校进行排序。

理想解法相关公式有以下：

$$c_j^* = \begin{cases} \max c_{ij}, & j\text{为效益型属性} \\ \min c_{ij}, & j\text{为成本型属性.} \end{cases} \tag{13}$$

- c_j^* 为第 j 列的正理想解（向量）。

$$c_j^0 = \begin{cases} \min c_{ij}, & j\text{为效益型属性} \\ \max c_{ij}, & j\text{为成本型属性.} \end{cases} \tag{14}$$

- c_j^0 为第 j 列的负理想解。

$$s_i^* = \sqrt{\sum_{j=1}^{n}(c_{ij} - c_j^*)^2} \tag{15}$$

- s_j^* 为第 i 行数据到正理想解的距离。

$$s_i^0 = \sqrt{\sum_{j=1}^{n}(c_{ij} - c_j^0)^2} \tag{16}$$

- s_j^0 为第 i 行数据到负理想解的距离。

$$f_i^* = \frac{s_i^0}{s_i^0 + s_i^*} \tag{17}$$

- f_i^* 为第 i 行数据的综合指标值（即第 i 所院校的综合指标值）。按照 f_i^* 的大小顺序就能得出方案的优劣次序，f_i^* 的值越大，对应院校越好。

下面以模型一为例，详细叙述整个建模求解过程，而对于后续模型本文将不再深入展开讨论，只列出最后结果以供比较。

表 5: 模型一标准化后的数据

人均专著	生师比	科研经费	逾期毕业率
0.000	1.000	0.0	0.545
0.125	0.547	0.2	0.750
0.375	0.278	0.4	1.000
1.000	0.000	1.0	0.000

表 6: 模型一标准化并加权后的数据

人均专著	生师比	科研经费	逾期毕业率
0.000	0.237	0.000	0.103
0.038	0.129	0.055	0.141
0.113	0.066	0.110	0.188
0.301	0.000	0.274	0.000

5.4.1 模型一

模型一的建立

模型一采用的标准化方法为 min-max 法，采用的加权方法为变异系数法。

由 min-max 法标准化公式 (1)（注：由本数据为 4 行 4 列，故以下公式中，皆有 $i = 1, 2, 3, 4, \quad j = 1, 2, 3, 4, \quad n = 4, \quad m = 4$）：

$$b_{ij} = \frac{a_{ij} - \min a_{ij}}{\max a_{ij} - \min a_{ij}}.$$

可得标准化后的数据如表 5 .

由变异系数法加权公式 (4)(5)(11)：

$$w_j = \frac{\sigma_j}{\overline{a}_j}$$

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(b_{ij} - \overline{b}_j)^2}$$

$$c_{ij} = w_j \cdot a_{ij}.$$

可得加权后的数据如表 6 .

表 7: 模型一理想解表

s^*	s^0	f^*
0.420	0.252	0.375
0.386	0.153	0.284
0.356	0.171	0.324
0.237	0.448	0.654

表 8: 模型二最终相关数据

	处理后的数据				理想解		
	人均专著	生师比	科研经费	逾期毕业率	s^*	s^0	f^*
院校一	0.000	0.246	0.00	0.134	0.383	0.271	0.414
院校二	0.032	0.135	0.05	0.185	0.371	0.160	0.301
院校三	0.097	0.068	0.10	0.246	0.375	0.155	0.292
院校四	0.257	0.000	0.25	0.000	0.246	0.435	0.639

模型一的求解

由表 5 可以看出当前模型的正理想解为：

```
(0.3, 0.24, 0.27, 0.0)
```

负理想解为：

```
(0.0, 0.0, 0.0, 0.19)
```

由正负理想解带入相应公式得到的数据如表 7 .

由表 6 可得知当前模型求解出的院校排名结果为：4>1>3>2

5.4.2 其余模型的建立与求解

模型二

模型二采用的标准化方法为 min-max 法，采用的加权方法为均方差法。

现直接给出模型二最终相关数据, 如表 8 .

由表 7 可得知当前模型求解出的院校排名结果为：4>1>2>3

模型三

模型三采用的标准化方法为 min-max 法，采用的加权方法为熵权法。

现直接给出模型三最终相关数据, 如表 9 .

由表 8 可得知当前模型求解出的院校排名结果为：4>1>3>2

表 9: 模型三最终相关数据

	处理后的数据				理想解		
	人均专著	生师比	科研经费	逾期毕业率	s^*	s^0	f^*
院校一	0.000	0.230	0.000	0.097	0.429	0.244	0.362
院校二	0.040	0.126	0.055	0.134	0.392	0.149	0.276
院校三	0.119	0.064	0.110	0.179	0.354	0.173	0.329
院校四	0.317	0.000	0.274	0.000	0.230	0.455	0.664

表 10: 模型四最终相关数据

	处理后的数据				理想解		
	人均专著	生师比	科研经费	逾期毕业率	s^*	s^0	f^*
院校一	-0.147	-0.162	-0.646	-0.008	1.69	0.130	0.071
院校二	-0.098	-0.027	-0.323	0.050	1.36	0.360	0.209
院校三	0.000	0.053	0.000	0.121	1.04	0.696	0.400
院校四	0.245	0.136	0.968	-0.164	0.00	1.711	1.000

模型四

模型四采用的标准化方法为 Z-score 法，采用的加权方法为变异系数法。

现直接给出模型四最终相关数据, 如表 10 .

由表 9 可得知当前模型求解出的院校排名结果为：4>3>2>1

模型五

模型五采用的标准化方法为 Z-score 法，采用的加权方法为均方差法。

现直接给出模型五最终相关数据, 如表 11 .

由表 10 可得知当前模型求解出的院校排名结果为：4>1>2>3

表 11: 模型五最终相关数据

	处理后的数据				理想解		
	人均专著	生师比	科研经费	逾期毕业率	s^*	s^0	f^*
院校一	-0.182	0.277	-0.200	-0.014	0.752	0.559	0.426
院校二	-0.122	0.046	-0.100	0.090	0.735	0.327	0.308
院校三	0.000	-0.091	0.000	0.217	0.760	0.306	0.287
院校四	0.304	-0.232	0.301	-0.292	0.509	0.864	0.630

表 12: 五个模型的排名结果

X.	第一名	第二名	第三名	第四名
模型一	4	1	3	2
模型二	4	1	2	3
模型三	4	1	3	2
模型四	4	3	2	1
模型五	4	1	2	3

表 13: 总得分排名

院校	得分
4	20
1	13
3	9
2	8

5.4.3 结论

五个模型分别求解的排名结果如表 12 .

假设第一至四名分别对应得分为 4、3、2、1，则它们的最终总得分排名如表 13 .

加上最初剔除的第五所院校（预先确定了其为第一名），本模型最终的排名结果为：5>4>1>3>2

六、模型的评价、改进与推广

6.1 模型的优点

1. 经由相关分析，人为剔除了与其他院校相差较大的第五所院校的数据，有一定道理。且该处理相当于消除极端值，减小了后面使用客观赋权法进行建模的误差；
2. 以一流大学的生师比数据为依据并拟合其分布，将原数据的生师比依概率赋值，较为合理；
3. 使用五个基于客观赋权法的模型结果进行平均得出了最后结果，较为有信服力。

6.2 模型的缺点

1. 人为剔除第五个院校的数据较为武断，缺乏更有力的资料支撑；
2. 由相关资料 [2]，小样本数据使用客观赋权法效果通常不理想，对于院校评价这样复杂的模型，需要结合专家的经验性看法才能得到较好的评估效果。

6.3 模型的推广与改进

本文对数据的指标进行了相关性分析，得出生师比、科研经费、人均专著间具有强相关关系。如有需求，可拟合相应回归方程，并根据其中一项的值来预测另外两项，如此便能由预测值与实际值的差距来对院校做额外的评估，比如评估该院校的相应指标是否处于正常范围等等。当然，一般来讲，仅基于四行数据的立论往往是不可靠的，但一定程度上也应具有其参考价值。

本文未包含主观赋权法的最大原因是在于找不到可信服的专家制定的权数，如有更多时间，可以在寻求专家权数方面多做努力。

参考文献

[1] 范哗. 大众化进程中的生师比与大学质量关系——世界一流大学生师比研究的启示 [J]. 教育发展研究,2012,32(23):8-15.DOI:10.14121/j.cnki.1008-3855.2012.23.005.

[2] 李健宁. 高等学校学科竞争力评价研究 [D]. 华东师范大学,2004.

[3] 郭显光. 熵值法及其在综合评价中的应用 [J]. 财贸研究,1994(06):56-60.DOI:10.19337/j.cnki.34-1093/f.1994.06.014.

[4] 赵静, 但琦. 数学建模与数学实验（第五版）[M]. 高等教育出版社, 北京, 2020.

附录

```{r}
#--------------------R 语言代码--------------------#

# 小数点位数定为 3 位
```

```r
options(digits = 3)
# 建立数据
df0 = data.frame(
    '人均专著' = c(0.1, 0.2, 0.4, 0.9, 1.2),
    '生师比' = c(5, 6, 7, 10, 2),
    '科研经费' = c(5000, 6000, 7000, 10000, 400),
    '逾期毕业率' = c(4.7, 5.6, 6.7, 2.3, 1.8)
)

# 相关系数图: 一
library(GGally)
ggpairs(df0) +
    theme_minimal() +
    theme(panel.grid = element_blank(),
        panel.border = element_rect(fill=NA),
        axis.text =  element_text(color='black'))

# 剔除第五行数据（第五所院校）
df0 = df0[-5, ]

# 相关系数图: 二
ggpairs(df0) +
    theme_minimal() +
    theme(panel.grid = element_blank(),
        panel.border = element_rect(fill=NA),
        axis.text =  element_text(color='black'))

# 回归拟合图
library(ggplot2)
ggplot(df0, aes(x=`生师比`, y=`人均专著`)) +
    geom_point() +
    geom_smooth(method = "lm")
```

```r
# 拟合伽马分布
ssb = c(10, rep(7, 2), rep(6, 3), rep(5, 7),
        rep(4, 9), rep(3, 15), rep(2, 18),
        rep(1, 12), rep(0.5, 7))
library(MASS)
fitt = fitdistr(ssb, "gamma")
df_ga = data.frame(
    x = rgamma(10000, shape = fitt$estimate[1],
               rate = fitt$estimate[2])
)
df_ssb = data.frame(
    x = ssb
)

# 拟合结果显著性检验
ks.test(jitter(ssb, 0.001), "pgamma", shape = 0.347, rate = 0.137)

# 伽马分布拟合图
library(latex2exp)
ggplot(mapping = aes(x)) +
    geom_density(data = df_ga, size = 1) +
    geom_histogram(data = df_ssb,
                   aes(y = ..density../1.75),
                   alpha = 0.6) +
    ylab(TeX(r'($p(x)$)'))

df0$生师比 = dgamma(c(5, 6, 7, 10),
                  shape = fitt$estimate[1],
                  rate = fitt$estimate[2])

# min-max 标准化法
```

```r
ax = function(df0_a){
    for (i in 1:4){
        df0_a[, i] = (df0_a[, i]-min(df0_a[, i]))/
            (max(df0_a[, i])-min(df0_a[, i]))
    }
    return(df0_a)
}

# Z-score 标准化法
bx = function(df0_b){
    avg = rep(NA, 4)
    sdd = rep(NA, 4)
    for (i in 1:4){
        avg[i] = mean(df0_b[, i])
        sdd[i] = sqrt(4/(4-1)*var(df0_b[, i]))
    }
    for (i in 1:4){
        df0_b[, i] = (df0_b[, i]-avg[i])/sdd[i]
    }
    return (df0_b)
}

# 变异系数法
xa = function(df0_xa){
   v = rep(NA, 4)
   for (i in 1:4){
     v[i] = sqrt((var(df0_xa[, i]))*4/(4-1))/mean(df0_xa[, i])
   }
   for (i in 1:4){
      df0_xa[, i] = df0_xa[, i]*v[i]/sum(v)
   }
  return (df0_xa)
```

```
}

# 均方差法
xb = function(df0_xb){
   s = rep(NA, 4)
   for (i in 1:4){
     s[i] = sqrt(4/(4-1)*var(df0_xb[, i]))
   }
   for (i in 1:4){df0_xb[, i] = df0_xb[, i]*s[i]/sum(s)}
   return (df0_xb)
}

# 熵权法
xc = function(df0_xc) {
    # 判定数据中是否有 0 或 1，有的话制造噪声
    chaos = function(l){
        l[l == 0] = 0.001
        l[l == 1] = 0.999
        return (l)
    }
    df0_xc = apply(df0_xc, 2, chaos)
    # 实现用熵权法计算各指标（列）的权重及各数据行的得分
    # s 返回各行（样本）得分，w 返回各列权重
    # 计算第 j 个指标下，第 i 个样本占该指标的比重 p(i,j)
    P = apply(df0_xc, 2, function(x) x / sum(x))

    # 计算第 j 个指标的熵值 e(j)
    e = as.vector(apply(P, 2, function(x) sum(x * log(x)) *
                              (-1/log(nrow(P)))))

    d = 1 - e          # 计算信息熵冗余度
    w = d / sum(d)     # 计算权重向量
```

```
    for (i in 1:4){
        df0_xc[, i] = df0_xc[, i] * w[i]
    }
    return (df0_xc)
}

# TOPSIS 法

## 正理想解与负理想解
good_bad = function(df){
    good = rep(NA, 4)
    for (i in 1:3){
        good[i] = max(df[, i])
    }
    good[4] = min(df[, 4])

    bad = rep(NA, 4)
    for (i in 1:3){
        bad[i] = min(df[, i])
    }
    bad[4] = max(df[, 4])

    return(list(good,bad))
}

## 正负理想解、综合指标数据表
over = function(df){
    gg = function(row){sqrt(sum((row - g)^2))}
    bb = function(row){sqrt(sum((row - b)^2))}
    df0_over = data.frame(
        s = rep(NA, 4),
```

```
        s0 = rep(NA, 4)
    )
    for (i in 1:4){
        df0_over$s[i] = gg(df[i, ])
        df0_over$s0[i] = bb(df[i, ])
    }
    df0_over$f = df0_over$s0/(df0_over$s + df0_over$s0)
    names(df0_over) = c("$s^*$", "$s^0$", "$f^*$")
    return (df0_over)
}
```