关于大学生学科竞赛赛事的评价分析与问题解决

摘　要

在大学时代，积极参与学科竞赛对大学生的发展具有多重意义。首先，竞赛是一种全面发展和自我实现的重要途径，可以提升学生的学业成绩和人格魅力。其次，通过竞赛，学生能够在学术上和专业领域取得优异成绩，为未来的发展奠定坚实的基础。最后，对于大学生而言，慎重地选择参加何种竞赛也是十分重要的，因为不同竞赛会给予不同的锻炼和机会，需要根据自身情况和发展目标进行选择。因此，对大学生各种竞赛赛事的评估和分析，对于他们的成长和发展具有重要意义。这需要综合考虑学生的兴趣爱好、专业特长、未来规划等因素，为其提供合适的竞赛机会，促进其全面发展和综合素质的提升。

针对第一问，由于原始数据样本数据太多造成计算难度较大，因此先系统选取了 14 个数据样本进行分析。选取赛事参加总人数、赛事举办届数、CNIK 数量、ALEXA 国内排名以及主办方权威性五个指标进行分析评价。使用**秩和比综合评价法**进行评估得到结论对选取的 14 个竞赛赛事进行了三等级的分级评价，分别为 A 级（3 个）、B 级（9 个）、C 级（2 个）。

针对第二问，首先对于给出的数据进行大范围的筛选，去除可能性较低进入正式目录的赛事，而后采用了**熵值法**进行总体评价分析，重新确立几个指标后进行进行评估，对数据进行处理与分析，用公式对极小型数据、中间型数据进行不同方式的处理使其成为极大型指标，正向处理后的指标对其进行标准化则得到处理后的数据。紧接着使用公式计算信息熵并得到权重，最终得出顺序排列的 12 个赛事排序，根据结果得到最有可能进入正式目录的赛事。

对于第三问，考虑分别探究三个因素对参赛人数的影响，考虑到三个因素不是简单的线性关系，所以先分别探讨三个因素与参赛人数的关系，建立回归方程。分别根据不同的限制条件建立三个具体的预测回归模型（具体模型公式见后文）。考虑到本文第一问从三个指标分别分析对参加人数的影响，构建的是线性回归模型和有理函数模型。第二问为发现三个指标之间的相互作用与内在关联，现重新构建模型，以寻找各因素间的相互影响和共同对因变量投入的贡献程度，所以选择构建**多项式非线性回归模型**。由于函数式较为复杂，故采用**粒子群算法**求函数最大值，最终得出结论，当每周理论培训时长为 290 分钟，实践培训时长为 260 分钟，投入 699 千元时，将有最多的参与人数，即有 45.216% 的人参加培训。最后第三小问需根据前文所述提出建议，具体建议见后文详述。

关键词： 秩和比综合评价法　熵值法　多项式非线性回归模型　粒子群算法

目　录

一、 问题重述

1.1 问题背景

在大学时代，参加各种学科竞赛是一种积极而富有意义的行为。这不仅是一种展示个人才华和专业知识的机会，更是一次锻炼和成长的过程。参加竞赛可以让学生在激烈的竞争中学会团队合作、解决问题的能力，并且通过与来自不同院校、不同背景的同学交流，拓展自己的视野，结交志同道合的朋友，建立起深厚的人际关系网络。因此，大学生积极参与学科竞赛，不仅是对自身能力的锻炼和提升，更是一种全面发展和自我实现的重要途径。通过竞赛，学生能够在学业上取得优异成绩，在人格上得到更深的塑造，为未来的发展奠定坚实的基础。由此，慎重地对大学生各种赛事的选择尤为重要，对大学生面临的各种竞赛赛事的评估和分析也成为值得深思的实际问题。

1.2 问题提出

本题给出了中国高等教育学会高校竞赛评估与管理体系研究专家工作组发布《2023全国普通高校大学生竞赛分析报告》中包含的13个普通本科院校大学生竞赛榜单、11个高职院校大学生竞赛榜单、3个省份大学生竞赛榜单。需根据各赛事的运营情况，确立评价标准和规则，建立分级条件并对数据进行处理分析，量化处理并建立模型，对各项赛事进行分级评价和排序。具体问题如下：

1. 对竞赛目录中的84项赛事进行筛选，选择合适的评价指标并对数据进行量化处理，最后对选取的赛事进行综合评价并分类。
2. 对进入观察目录的34项赛事进行评价指标的选取，并选择评价方法进行综合评价分析，预测出最有可能进入竞赛目录的赛事。
3. 根据设置的不同问卷数据，要求建立参赛的人数与每周理论和实践培训以及投入之间的数学模型。并针对建立的模型，推算讨论得到使得参加学生比例呈现最高水平的最佳的培训时间和投入。最后根据建模以及讨论的结果从综合应用价值和改进方式等方面给学校相关部门提出对应的建议和意见。

二、 问题分析

2.1 问题一分析

针对第一问，由于原始数据样本数据太多造成计算难度较大，因此先系统选取了14个数据样本进行分析。选取合适的指标进行分析评价。使用秩和比综合评价法进行评估得到结论对选取的14种竞赛赛事进行了三等级的分级评价。

2.2 问题二分析

本题要求给出赛事的科学评价方法，从而对其综合评价并排序，分析题目知属于多指标决策分析的问题，由此选择熵值法这一常用于多指标决策分析的数学方法，其核心原理在于，对于数据集中的每个指标，通过计算其分布的不确定性，来衡量其在综合决策中的重要性。这种方法既考虑到了指标之间的关联性和权重分配，又具有较好的稳定性和可操作性。先对数据采用正向处理，后对其标准化，计算信息熵以确定权重，可使决策者可以更好地比较和评估不同选项，并根据计算出来的权重，对评估对象进行排名，以达到本题解题目标。具体分析流程如下：

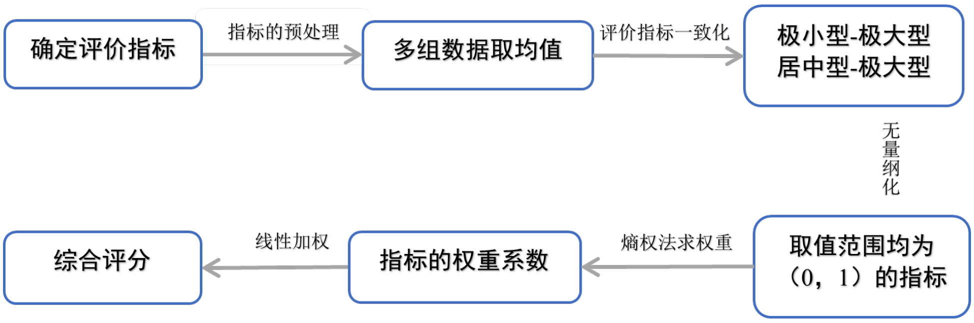


图 1　问题二分析流程图

2.3 问题三分析

针对第三问，分别探究三个因素对参赛人数的影响，考虑到三个因素不是简单的线性关系，所以先分别探讨三个因素与参赛人数的关系，建立回归方程。分别根据不同的限制条件建立三个具体的预测回归模型（具体模型公式见后文）。考虑到本文第一问从三个指标分别分析对参加人数的影响，构建的是线性回归模型和有理函数模型。第二问为发现三个指标之间的相互作用与内在关联，现重新构建模型，以寻找各因素间的相互影响和共同对因变量投入的贡献程度，所以选择构建多项式非线性回归模型。

三、 模型假设

在建模的过程中，为简化问题且方便建模，我们在不影响模型的可靠性和有效性的前提下，做出以下假设：

（1）假设竞赛表现优异的学生往往也在学业上有较好的表现。因此，竞赛成绩与学生的学业成绩之间存在一定的正相关关系。

（2）假设积极参与竞赛可以促进学生的综合素质提升，包括但不限于专业知识、团队合作能力、创新思维、沟通能力等方面。通过参与竞赛，学生可以在实践中不断提升自身的能力和素质水平。

（3）假设参与竞赛并获得一定成绩的学生，其未来的学习和职业发展可能会受到积极影响。竞赛经验可以为学生的升学、就业和科研等方面提供额外的加分项，增强其竞争力。

（4）假设参赛学生投入更多的时间和精力准备竞赛，往往能取得更好的成绩。因此，竞赛成绩与学生准备竞赛的时间投入之间存在一定的正相关关系。

四、 符号说明

符号	说明
a_0	线性回归方程常数项
a_1	线性回归方程一次项
a_2	线性回归方程二次项
b_1	非线性回归方程分子一次项
b_2	非线性回归方程分子常数项
b_3	非线性回归方程分母常数项

五、 模型的建立与求解

5.1 第一小问模型的建立与求解

5.1.1 基于秩和比综合评价模型的建立

针对本问使用秩和比综合评价法（RSR法）来建立在不排序的情况下对数据进行分级的模型[1]。RSR法的基本思想是在一个n行（n个评价对象）p列（p个评价指标）矩阵中，通过秩转换，获得一个无量纲的统计量RSR值，再以RSR值对评价对象的优劣进行分档。

首先构建模型，步骤如下：

Step1：选取评价对象

对于84个给出的大学生赛事直接进行评价会产生较多的工作量，故根据重要性和可评性以及评估的实用性选取了14个大学生赛事进行整体评估，其分别为中国高校计算机大赛、全国大学生物流设计大赛、全国大学生结构设计竞赛、全国大学生节能减排社会实践与科技竞赛、全国大学生机械创新设计大赛、全国大学生机器人大赛、全国大学生化工设计竞赛、全国大学生广告艺术大赛、全国大学生电子设计竞赛、全国大学生电子商务“创新、创意及创业”挑战赛、两岸新锐设计竞赛·华灿奖、ACM-ICPC 国际大学生程序设计竞赛、“西门子杯”中国智能制造挑战赛以及“挑战杯”全国大学生课外学术科技作品竞赛。为方便接下来的计算和表达，按以上给出的顺序依次将这14个比赛编号1~14。由此选择出数量较少的赛事，方便数据的收集和分析。

Step2：选取评价指标并同趋化处理数据

影响对大学生竞赛赛事评价的指标多种多样，从各方面皆有涉及。通过查阅文献我们进行合理的指标选取，最终保留赛事参加总人数、赛事举办届数、CNIK 数量、ALEXA

国内排名以及主办方权威性共计四项指标来进行评价，其五项指标分别展现了赛事的参与度、办赛经验、赛事研究价值、赛事热度以及赛事的权威性能，足以公平有效地对各项大学生赛事评价和给出评级。

Step3：确认指标权重

对数据的处理包括对其进行清洗和预处理，涵盖去除缺失数据、异常值的处理等。为简化处理，对于主办方的设定采取设置不同大小的数值的方式，更加重要的主办方选取更大的数字，例如多部委（如教育部）主办的选择更大的数字，而协会和企业主办的则选用较小的数字。

对于数据进行同趋化处理与量纲问题的考虑，并采用熵权法确认各指标权重，计算出的指标权重如下表：

表 1　各项指标权重

熵权法			
项	信息熵值d	信息效用熵值e	权重%
参赛人数	0.295	0.705	51.256
CNIK 数	0.712	0.288	20.942
项	信息熵值d	信息效用熵值e	权重%
举办届数	0.843	0.157	11.419
ALEXA 国内排名	0.921	0.079	5.764
主办方权威度	0.854	0.146	10.619

根据上表展示的由熵权法得出的权重计算结果，对各指标权重进行分析如下：熵权法的权重计算结果显示，参赛人数的权重为51.256%、CNIK 数的权重为20.942%、举办届数的权重为11.419%、ALEXA 国内排名的权重为5.764%、主办方权威度的权重为10.619%，其中指标权重最大值为参赛人数（51.256%），最小值为 ALEXA 国内排名（5.764%）。

Step4：秩值的计算

根据每一个具体的评价指标按其指标值的大小进行排序，得到秩次，用秩次 来代替原来的评价指标值，根据编秩结果建立各指标的秩次数据矩阵。计算秩的方法大致分为两种，为整秩法和非整秩法。整秩法即将 个评价对象的 个评价指标排列成 行 列的原始数据表。编出每个指标各评价对象的秩，其中效益型指标从小到大编秩，成本型指标从大到小编秩，同一指标数据相同者编平均秩。得到秩矩阵 ；而非整秩法用类似于线性插值的方式对指标值进行编秩，以改进 法编秩方法的不足，所编秩次与原指标值之间存在定量的线性对应关系，从而克服了 法秩次化时易损失原指标值定量信息的缺点。

因此，本题选择采用整秩法将每一个具体的考评指标数值按高优指标从小到大进行编秩，得到秩次 ，再用秩次 来代替原来的考评指标值。根据编秩结果建立各考评指标的秩次数据矩阵：

$$R=\begin{pmatrix} R_{1,1} & \cdots & R_{1,5} \\ \vdots & \ddots & \vdots \\ R_{19,1} & \cdots & R_{19,5} \end{pmatrix}$$

Step5：RSR值的计算

利用 Step2 的秩次，计算得到 值：

$$\mathrm{RSR}_n=\frac{1}{17\times 3}\sum_{p=1}^{3}R_{n,p}$$

Step6：分布表格的列写与Probit值的求得 RSR的分布是指用概率单位 表达的值特定的累计频率。其方法如下：

1. 将RSR值按照从小到大的顺序排列；
2. 列出各组频数f；
3. 计算各组累计频数；
4. 确定各组RSR的秩次R及平均秩次$\overline{R}$；
5. 计算向下累计频率$\left(\overline{R}-/n\right)\times 100\%$,最后一项用修正；
6. 根据累计频率，查询“百分数与概率单位对照表”，求其所对应概率单位Probit值。

值得注意的是，系统在编秩过程中进行的是同向趋势化处理，即将负向指标（成本型指标）转化成正向指标（效益型指标），统一对所有指标进行从小到大编秩。

Step7：直线回归方程的计算与对应RSR估计值的拟合

利用最小二乘估计法，以概率单位值Probit为自变量，以RSR值为因变量，计算线性回归方程。得出如下形式的结果：

$$\widehat{RSR}=a+b\times \mathrm{Probit} \tag{1}$$

Step8：依据拟合RSR值的排序与分级

按照回归方程推算得到的RSR估计值对考评对象进行分档排序。分档数量定为3档，分别对应优、中、差。

5.1.2 基于综合评价模型的求解

对本模型进行求解得出表格 2、表格 3，表格 2 给出了考评指标数据表和RSR的计算结果，表格 3 给出了RSR值分布。

表 2

索引	X_1:参赛人数	R_1:参赛人数	···	X_5:主办方权威度	R_5:主办方权威度	RSR	排名
1	0.01	1.07	···	0.50	7.50	0.19	10
2	0.00	1.00	···	0.00	1.00	0.08	14

索引	X_1:参赛人数	R_1:参赛人数	···	X_5:主办方权威度	R_5:主办方权威度	RSR	排名
3	0.00	1.03	···	0.00	1.00	0.15	12
4	0.01	1.18	···	0.50	7.50	0.19	9
5	0.01	1.11	···	0.00	1.00	0.24	4
6	0.00	1.02	···	0.50	7.50	0.19	8
7	0.01	1.08	···	0.50	7.50	0.23	5
8	0.02	1.28	···	0.50	7.50	0.19	7
9	0.02	1.31	···	1.00	14.00	0.42	3
10	0.00	1.03	···	0.50	7.50	0.15	13
11	0.16	3.10	···	0.00	1.00	0.21	6
12	0.00	1.00	···	0.50	7.50	0.47	2
13	0.01	1.08	···	0.50	7.50	0.17	11
14	1.00	14.00	···	1.00	14.00	0.84	1

表 3

RSR	频数	累计频数$\sum f$	评价秩数	评价秩数$/n \times 100\%$	Probit
0.085	1	1	1	7.143	3.535
0.150	1	2	2	14.286	3.932
0.152	1	3	3	21.429	4.208
0.166	1	4	4	28.571	4.434
0.189	1	5	5	35.714	4.634
0.193	1	6	6	42.857	4.820
0.194	1	7	7	50.000	5.000
0.195	1	8	8	57.143	5.180
0.207	1	9	9	64.286	5.366
0.227	1	10	10	71.429	5.566
0.235	1	11	11	78.571	5.792
0.418	1	12	12	85.714	6.068
0.472	1	13	13	92.857	6.465
0.843	1	14	14	98.214	7.100

运用 SPSS 软件求解得 Step6 中线性回归方程式(1)中的常量$a = -0.601$和系数$b = 0.168$，故可以得到对应的线性回归方程。

表 4

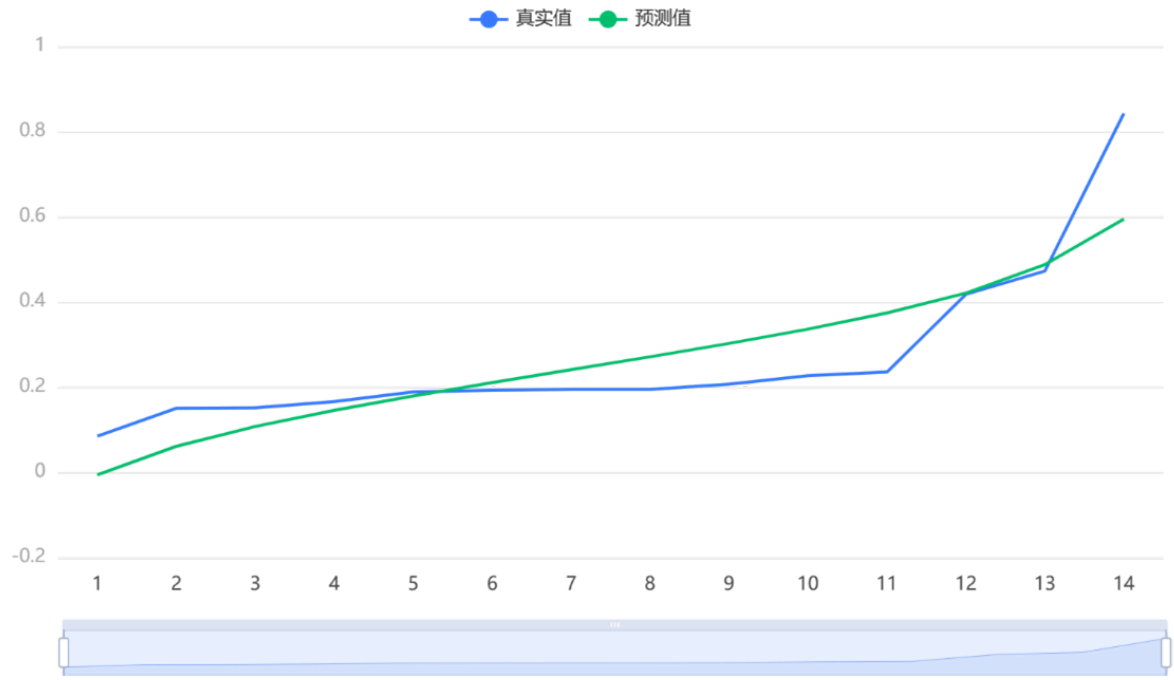
线性回归分析结果$b = 14$									
	非标准化系数		标准化系数	t	P	VIF	R^2	调整R^2	F
	B	标准错误	Beta						
常数	-0.601	0.151	-	-3.972	$0.002***$	-	0.739	0.717	

Probit	0.168	0.029	0.86	5.827	0.002***	1	$F=33.959$ $P=0.000***$
因变量：RSR							
注：***、**、*分别代表1%、5%、10%的显著性水平							

表 4 展示了本次模型的分析结果，包括模型的标准化系数、t值、VIF值、R^2、调整R^2等，用于模型的检验，并分析模型的公式。其中R^2代表曲线回归的拟合程度，越接近1效果越好；VIF值代表多重共线性严重程度，用于检验模型是否呈现共线性，即解释变量间存在高度相关的关系（VIF应小于10或者5，严格为5）。若VIF出现inf，则说明VIF值无穷大，建议检查共线性，或者使用岭回归。线性回归模型要求总体回归系数不为0，即变量之间存在回归关系，根据 F 检验结果对模型进行检验。

从 F 检验的结果分析可以得到，显著性P值为$0.000***$，水平呈现显著性，拒绝了回归系数为 0 的原假设，同时模型的拟合优度R^2为0.717，模型表现较为优秀，因此模型基本满足要求。对于变量共线性表现，VIF全部小于10，因此模型没有多重共线性问题，模型构建良好。

由此输出拟合效果图，展示本次模型的原始数据图、模型拟合值、模型预测值如下：



通过回归方程计算得到的RSR估计值分档排序结果如表 5 所示。

表 5　分档排序结果

分档排序临界值表格			
档次	百分比临界值	Probit	RSR临界值（含拟合值）

分档排序临界值表格			
第 1 档	< 15.866	< 4	< 0.0724
第 2 档	15.866~	4~	0.0724~
第 3 档	84.134~	6~	0.4092~

本步骤目的在于得到分档排序临界值表格，尤其是Probit临界值对应的 RSR 临界值（拟合值）。百分位数临界值和RSR临界值根据分档水平数量而变化，该两项是固定值且完全一一对应；而表格中RSR临界(拟合值)则是根据Probit临界值代入回归模型计算得到。

接着为了将数据按照秩的各种情况映射到正态分布曲线上，结合正态分布的相关划分方法进行分档，需建立分档等级结果。先通过RSR拟合值，以及上一表格中的RSR临界(拟合值)进行区间比较，进而得到分档等级水平。分档等级Level数字越大表示等级水平越高，即效应越好。

由此得出各项竞赛分档等级结果汇总如下：

表 6　分档结果

名称	RSR排名	Probit	RSR拟合值	分档等级
全国大学生电子设计竞赛	3	6.465233792685523	0.4875824299071976	A
ACM-ICPC 国际大学生程序设计竞赛	2	6.067570523878141	0.4206049020488354	A
“挑战杯”全国大学生课外学术科技作品竞赛	1	7.100165492844469	0.5945225447513265	A
中国高校计算机大赛	10	6.067570523878141	0.4206049020488354	A
全国大学生结构设计竞赛	12	4.208361392256625	0.10746249948004671	B
全国大学生节能减排社会实践与科技竞赛	9	4.819987630207295	0.2104773278651365	B
全国大学生机械创新设计大赛	4	5.791638607743375	0.3741303114854323	B
全国大学生机器人大赛	8	5	0.2407964054827395	B
全国大学生化工设计竞赛	5	5.565948821932863	0.3361178896648158	B
全国大学生广告艺术大赛	7	5.1800123697927045	0.2711154831003424	B
两岸新锐设计竞赛·华灿奖	6	5.36610635680057	0.3024588737948707	B
“西门子杯”中国智能制造挑战赛	11	4.434051178067137	0.1454749213006632	B
全国大学生物流设计大赛	14	3.534766207314477	−0.005989618941718677	C
全国大学生电子商务“创新、创意及创业”挑战赛	13	3.9324294761218583	0.06098790891664352	C

综上即完成题设要求，对选取的 14 种竞赛赛事进行了分级评价，由上表所示分成了 A 级（三个）、B 级（9 个）、C 级（2 个）。

5.2 第二小问模型的建立与求解

5.2.1 基于综合评价模型的建立

本题要求未进入目录的34项赛事进行评价并排序，用于多准则决策分析方法，选择熵值法作为本题模型。

Step1：赛事的筛选

由于本题意欲寻找最有可能进入目录的赛事，故可先大幅进行筛选。经完成率和获奖率的筛选最终选择14个赛事作为最终的评价样本。

Step2：数据的正向处理

本题选择举办届数、参赛院校数、在观察目录中的排名以及主办方权威度四个方面作为指标。

目标是将指标数值表示为极大型，即数值越大代表指标越优秀。

分析举办届数为中间型指标，使用以下公式处理：

$$M = \max\{|x_i - x_{\text{best}}|\}$$

$$\tilde{x} = 1 - \frac{|x_i - x_{\text{best}}|}{M}$$

对于观察目录中的排名指标而言，其为极小型指标，应用以下公式将其化为极大型：

$$x^* = \frac{1}{x}(x > 0)$$

Step3：数据的标准化

数据标准化是一种将具有不同尺度和单位的数据转换为具有相同尺度和单位的数值的数据处理方法，通常用于多个维度数据的比较和分析。本题选择 min-max 标准化法进行数据标准化，公式如下：

$$\frac{x - \min}{\max - \min}$$

标准化后的数据一般在$[0,1]$或$[-1,1]$的范围内，可以根据需要进行调整。

Step4：熵值的原理

熵值法的核心原理在于，对于数据集中的每个指标，通过计算其分布的不确定性，来衡量其在综合决策中的重要性。这种方法既考虑到了指标之间的关联性和权重分配，又具有较好的稳定性和可操作性，是一种有效的多指标决策分析方法。

计算信息熵采用如下公式：

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

其中，信息熵越大，表示信息越不确定，呈现出的分布越均匀。通过计算所有指标的信息熵，可以得到每个指标的权重。

5.2.2 基于综合评价模型的求解、评价过程与结果

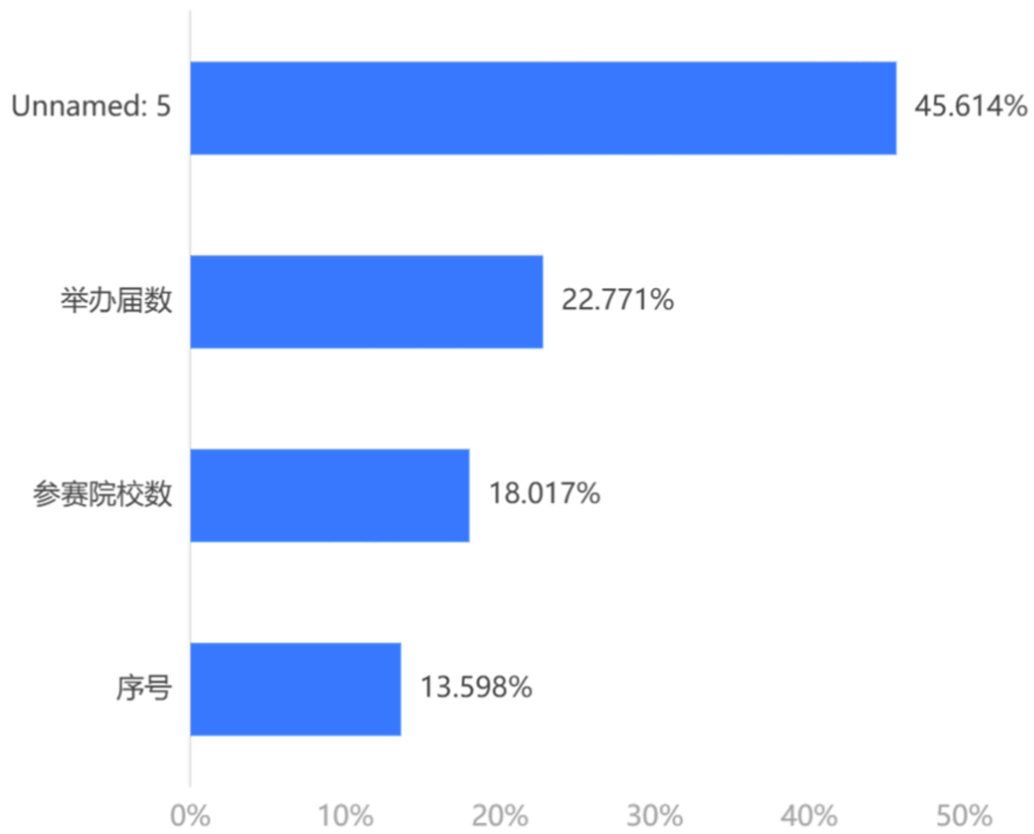
首先根据权重计算结果对各指标的权重进行分析。权重计算结果如下：

表 7

熵权法			
项	信息熵值e	信息效用值d	权重(%)
举办届数	0.852	0.148	22.771
参赛院校数	0.883	0.117	18.017
观察目录排名	0.912	0.088	13.598
主办方权威度	0.703	0.297	45.614

表 7 展示了熵值法的权重计算结果，根据结果对各个指标的权重进行分析。熵值法的权重计算结果显示，举办届数的权重为22.771%、参赛院校数的权重为 18.017%、观察目录排名的权重为13.598%、主办方权威度的权重为45.614%。

根据以上数据绘制指标重要度直方图，以直方图形式展示了指标的重要度排序（降序）如下所示：



根据权重输出得到综合得分，实现排序如表 8 所示：

表 8

行索引	综合评价	排名
1	0.17805508971552064	11
2	0.6258971475098188	4
3	0.0770851230799413	12
4	0.29324330502927876	8
5	0.23039785912635463	10
6	0.3057240998089282	7
7	0.5439566561694258	5
8	0.9167211059094001	1
9	0.4999018317876996	6
10	0.25229332695910317	9
11	0.9064190898634514	2
12	0.7201096205873163	3

5.3 第三问模型的建立与求解

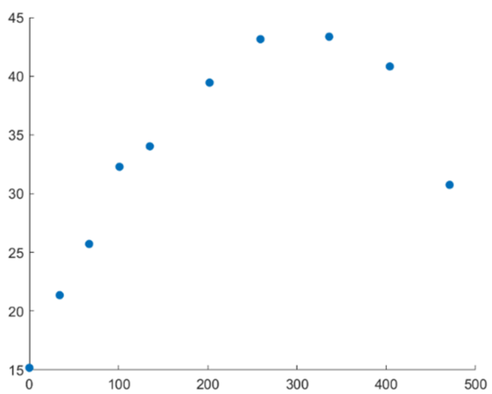
5.3.1 基于预测模型建立的问题分析

为了从应用价值和如何改进方面给某校相关部门提出建议，本论文拟打算分别探究三个因素对参赛人数的影响，考虑到三个因素不是简单的线性关系，所以先分别探讨三个因素与参赛人数的关系，建立回归方程。

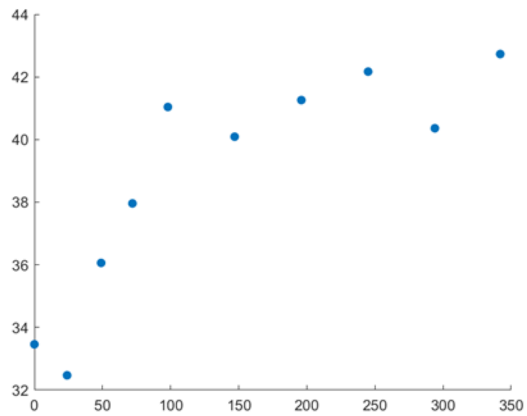
5.3.2 模型构建与分析

1 参赛人数与每周理论培训时间，每周实践培训时间的关系

1.1 绘制散点图

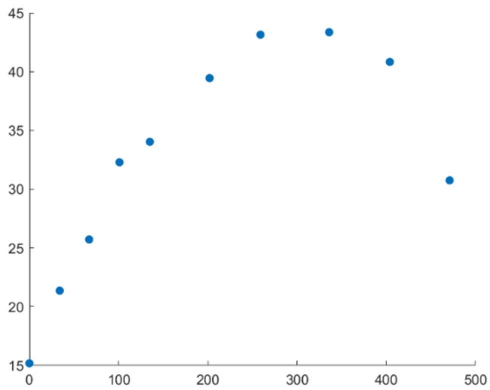


参赛人数-每周理论培训时间

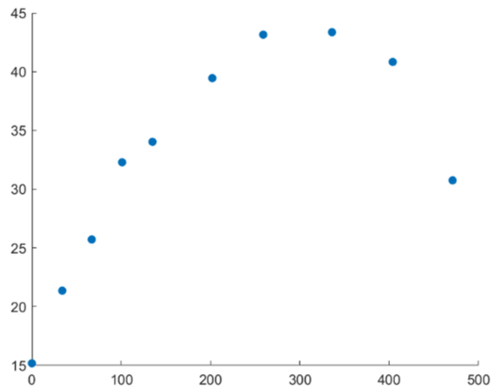


参赛人数-每周实践培训时间

在绘制参赛人数-每周实践培训时间的散点图时发现第二个数据异常，在进行线性插值后重新得到散点图如下图所示：



参赛人数-每周理论培训时间



参赛人数-每周实践培训时间

1.2 求解线性回归模型

1.3 模型参数求解与显著性检验

求解回归模型常用 matlab 中的 regress 函数，其中:

- a为回归系数，升幂排列
- aint为回归系数的置信区间，默认95%
- r为残差，即理论值与测量值的差
- rint为残差的95%的置信区间
- stats模型检验参数$(R^2, F, p, \mathrm{sig}^2)$（$R^2$表示方程因变量与自变量之间的相关性检验，$R$也称为可决系数；$F$，$p$是方程显著性检验；$\mathrm{sig}^2$是$\sigma^2$的估计值。）[2]最后得到参赛人数与每周理论培训时间的关系：

$$y_1 = 14.7535 + 1.1969x_1 - 0.0003x_1^2$$

参赛人数与每周实践培训时间的关系：

$$y_2 = 33.5964 + 0.0599x_2 - 0.0001x_2^2$$

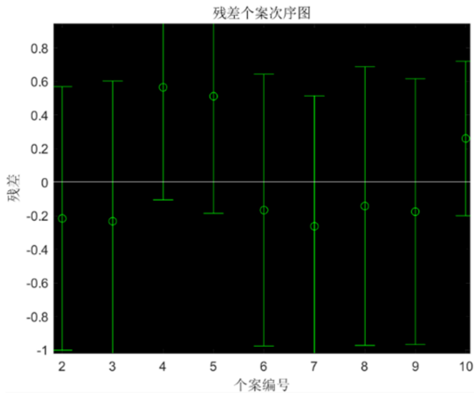
代码中`[ones(length(x1),1),x1',(x1.^2)']`矩阵，形式是：

$$\begin{vmatrix} 1 & x_1 & \dots & x_1^p \\ 1 & x_2 & \dots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^p \end{vmatrix}$$

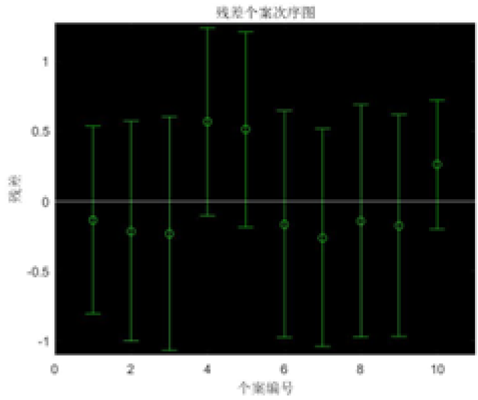
回归系数计算如下：

a_0	a_1	a_2
15.3716	0.1824	−0.0003

a_0	a_1	a_2
33.5964	0.0599	−0.0001



参赛人数-每周理论培训时间



参赛人数-每周实践培训时间

回归系数的置信区间一般不应包括0；若包括0，说明该系数不显著（100次实验，95次以上离0非常近，贡献不大），而由上表可以发现参赛人数与每周理论培训时间的回归方程的置信区间不含0，说明每周理论培训时间对参赛人数影响效果显著，同理，每周实践培训时间的回归方程的置信区间也不含0，说明每周理论培训时间对参赛人数影响效果也较为显著。计算得显著性水平（p-value）分别均约为0.000，说明方程显著性较强，说明曲线拟合效果较好。

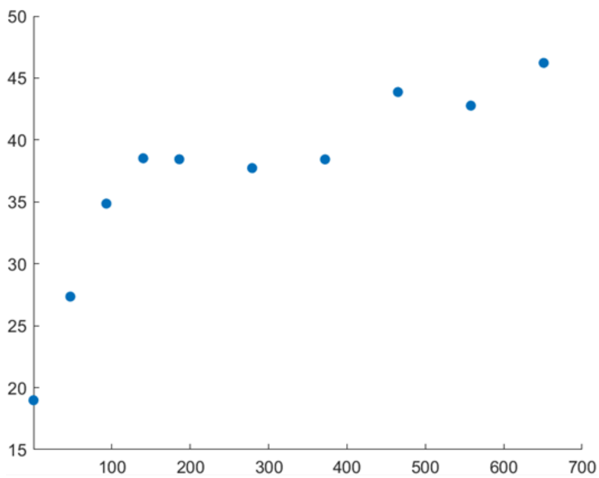
所以，回归方程可信度较高，参赛人数-每周理论培训时间和参赛人数-每周实践培训时间关系分别如下。

$$y_1 = 14.7535 + 1.1969x_1 - 0.0003x_1^2$$

$$y_2 = 33.5964 + 0.0599x_2 - 0.0001x_2^2$$

2 参赛人数与投入的关系

2.1 绘制散点图



根据散点图发现参赛人数与投入之间的关系与参赛人数与先前两个自变量的关系不同，不能单一地用线性回归方程拟合，根据实际问题分析得：随着投入的增大参赛人

数不可能一直增大，一定会达到一个极值，经过推断尝试发现用有理函数拟合效果最好。

2.2 非线性回归模型

$$y_3 = \frac{b_1 x_3 + b_2}{x_3 + b_3}$$

2.3 模型参数求解与显著性检验

计算得参赛人数与投入的关系：

$$y_3 = \frac{46.98x_3 + 1613}{x_3 + 85.62}$$

回归系数：

b_0	b_1	b_2
46.98	1613	85.62

上表可以发现参赛人数与投入·的回归方程的置信区间不含0，说明投入对参赛人数影响效果显著。由计算得显著性水平（p-value）均为0.000，说明方程显著性较强，说明曲线拟合效果较好。所以，回归方程可信度较高，参赛人数与投入的关系如下：

$$y_3 = \frac{b_1 x_3 + b_2}{x_3 + b_3}$$

综上，理论培训时间、实践培训时间、投入分别与参加人数的对应关系为：

$$y_1 = 14.7535 + 1.1969x_1 - 0.0003x_1^2$$

$$y_2 = 33.5964 + 0.0599x_2 - 0.0001x_2^2$$

$$y_3 = \frac{46.98x_3 + 1613}{x_3 + 85.62}$$

5.4 第三问第二小问模型的构建和求解

5.4.1 模型说明

考虑到本文第一问从三个指标分别分析对参加人数的影响，构建的是线性回归模型和有理函数模型。第二问为发现三个指标之间的相互作用与内在关联，现重新构建模型，以寻找各因素间的相互影响和共同对因变量投入的贡献程度，所以选择构建多项式非线性回归模型。

5.4.2 指标选取

本模型采用针对每周理论培训时间、每周实践培训时间以及投入的关系设置的不同问卷数据，并对其进行处理，作为模型的主要驱动因素。由于在该问卷中控制其它变量均为第七个水平上，为方便数理统计，分析每个因素对参加人数影响，可将每组数据用

第七项数据减去每项的参赛人数（即用43.15减去每组数据）便于分析。下面将针对这三个主要驱动因素分别对参赛人数的影响进行分析。

5.4.3 回归拟合

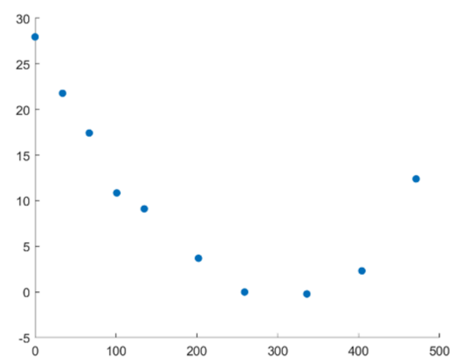
- 每周理论培训时间（分钟）

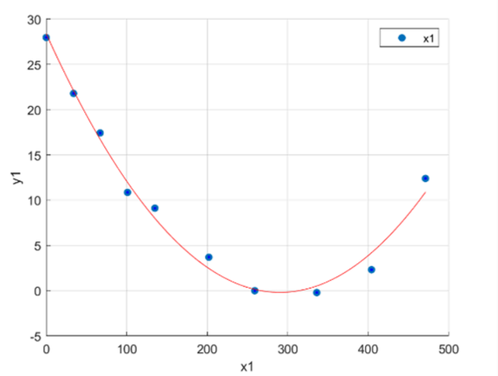
通过绘制出该数据的散点图，可猜想该函数为二次曲线，故建立如下模型：

$$y_1 = a_0 + a_1 x_1 + a_2 x_1^2$$

通过 `matlab` 中的 `fit` 函数进行回归拟合，可得到如下结果：

	未标准化系数		标准化系数		显著性
	B	标准化错误	Beta	t	
X	-.197	.010	−3.334	−19.949	<.001
X^2	.000	.000	2.752	16.466	<.001
(常量)	28.396	.887		32.018	<.001





从表中X项未标准化系数B为−0.197，标准化系数(Beta)为−3.334，t值为−19.949，显著性(p-value)小于0.001。这说明X项与因变量呈负相关，并且有显著影响，而 项未标准化系数B为0.0，但标准化系数(B)为2.752，t值为16.466，显著性(p-value)小于0.001。这表明 的平方项对因变量有显著的正相关影响。对于常数项未标准化系数B为28.396，t值为32.018，显著性(p-value)小于0.01，表明截距在统计上也是非常显著的。

通过相关系数计算公式可求得该回归曲线，可知该模型有较好的拟合效果，故猜想为二次曲线合理，即回归方程为:

$$y_1 = 28.396 - 0.1969 x_1 + 0.0003392 x_1^2$$

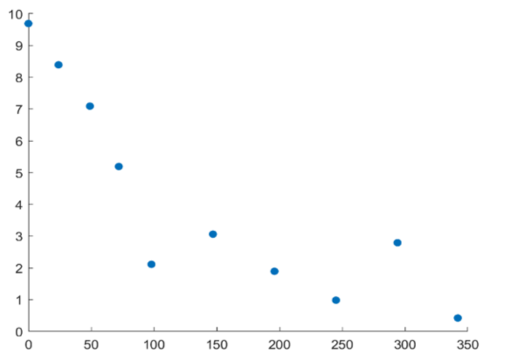
- 每周实践培训时间（分钟）

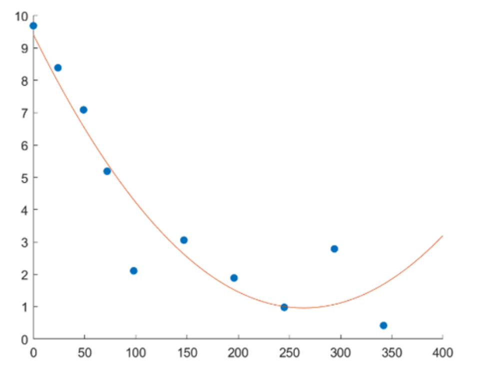
通过绘制出该数据的散点图，猜想该函数也为二次曲线，故建立如下模型：

$$y_2 = a_0 + a_1 x_2 + a_2 x_2^2$$

通过 `matlab` 中的 `polyfit` 函数进行回归拟合，可得到如下结果：

	未标准化系数		标准化系数		显著性
	B	标准化错误	Beta	t	
X	-.064	.013	2.336	4.887	.002
X^2	.000	.000	−1.535	−3.212	.015
(常量)	33.732	.849		39.711	<.001





通过相同方法可算相关系数$R^2 = 0.9826$，同时检验该回归模型的t值与显著性(p-value)，说明该模型有较好的拟合效果，故：

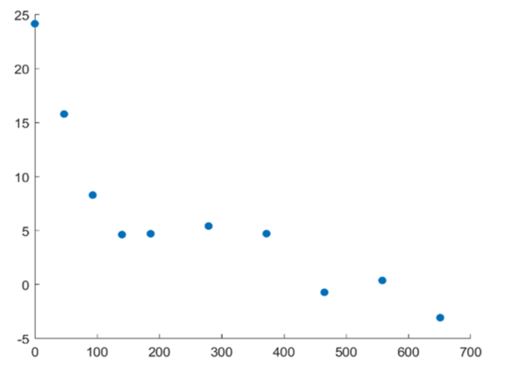
$$y_2 = 9.479 - 0.064x_2 + 0.0001212x_2^2$$

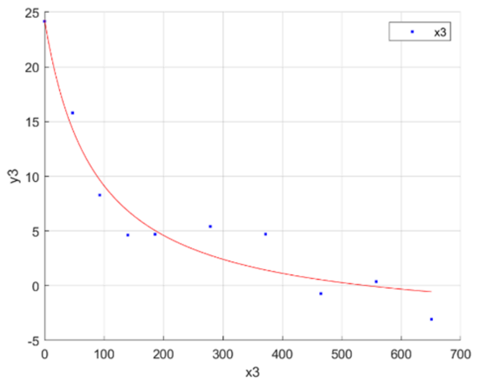
- 投入（千元）

通过绘制出该数据的散点图分布并实际结合投入与参加人数的关系，猜想该函数为有理函数，故建立如下模型：

$$y_3 = \frac{b_1 x + b_2}{x + b_3}$$

现通过 `matlab` 中的 `fit` 函数对该曲线进行拟合，所得曲线拟合情况如下：





与此同时，可以计算得到：SSE $= 34.9609$，调整后$R^2 = 0.9412$，故可知该曲线拟合效果较好，故：

$$y_3=\frac{-3.829x+2082}{x+85.62}$$

根据上述三个回归模型，分别得到每周理论培训时间、每周实践培训时间、投入对愿意参加人数影响的关系函数（以理论培训时间259，每周实践培训时间196，投入372，参与人数43.15%为标准），现将三条拟合曲线化成参加人数y与每周理论培训时间x_1、每周实践培训时间x_2、投入x_3的一个多元函数：

$$y=43.15-[y_1(x_1)+y_2(x_2)+y_3(x_3)]$$

即：

$$y=-0.0003392x_1^2-0.0001212x_2^2+0.1069x_1+0.064x_2+\frac{3.829x_3-2082}{x_3+85.62}+73361$$

根据公式可求得回归平方和：$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y_i}-\overline{y})^2$，总平方和：$\mathrm{SST}=\sum_{i=1}^{n}(y_i-\overline{y})^2$，和方差$\mathrm{SSE}=\sum_{i=1}^{n}(y_i-\hat{y_i})^2$，可计算出相关系数$R^2=1-\frac{\mathrm{SSE}}{\mathrm{SST}}=0.9602$，接近1，故原数据的总变异可以由拟合值的变异解释。

F 检验：

假设$H_0:\beta_1=0,H_1:\beta_1\neq 0$，其中：$\beta_1$为$y_i$关于$x_i$的总体回归系数。

现给出统计量F计算公式：$F=\frac{\mathrm{SSR}/(m-1)}{\mathrm{SSE}/(n-m)}=\frac{\mathrm{MSR}}{\mathrm{MSE}}$，其中，$n$为样本数量，$m$为自变量数量，即$n-m$为该模型的自由度。根据程序计算得到$F=348.8512$，通过查$F$检验临界表可知$F_{\alpha=0.05}=19.458>F$，故否定$H_0$假设，接受$H_1$假设，通过$F$检验。

将该回归模型代入原数据验证，可得到参与人数的实际值与拟合值之间的相对误差平均值在3%以内，表明对于拟合值和实际效果来说，拟合效果十分显著，证明该拟合出的参与人数方程通过检验。

5.4.4 计算最佳培训时间与投入

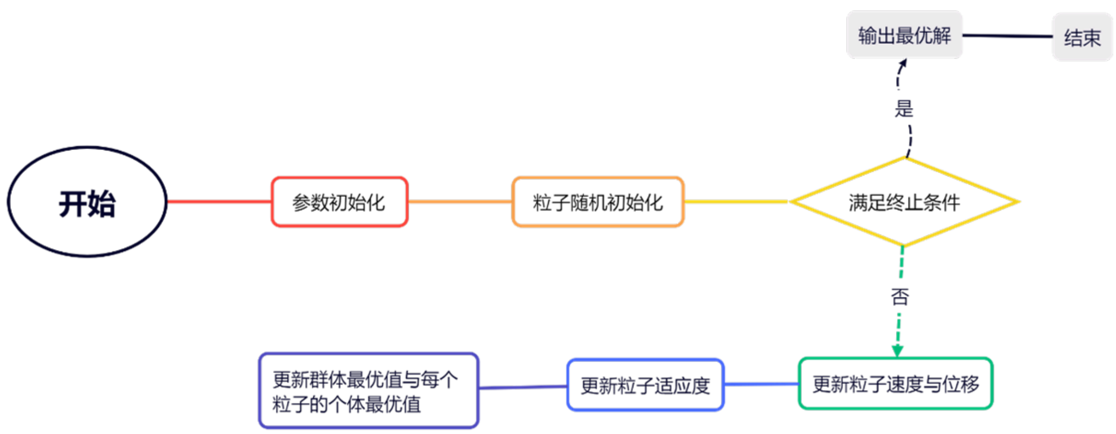
- 目标

$$y=-0.0003392x_1^2-0.0001212x_2^2+0.1069x_1+0.064x_2+\frac{3.829x_3-2082}{x_3+85.62}+73361$$

根据拟合的函数表达式，即求该函数取得最大值时的自变量的值。由于该函数存在多个变量，且为非线性函数，传统的Newton法等计算量大，求导困难且收敛性与结果很受初始值的影响，故现在避开传统数学方法，使用粒子群算法（PSO）求得该函数在一定区间内的最大值问题，该方法收敛快，且简单易实现。

- 粒子群算法（PSO：Particle swarm optimization）

粒子群算法源于鸟群捕食的行为研究，其基本思路是通过群体与个体之间的协作和信息共享来寻求最优解。[3]粒子仅具有两个属性：速度与位置。算法首先将初始化一组粒子群，每个粒子均具有一个初始位置和速度，根据适应度函数计算得到适应度函数值，将其标记成当前个体历史最优解，然后各个粒子之间通过协作共享信息得到全局历史最优解，并进行迭代得到新的速度和位置，然后计算粒子新的适应度函数值，如果粒子新的适应度函数值大于自己之前的个体历史最优解，那么这个新的适应度函数值成为该粒子新的个体历史最优解，所有粒子的个体历史最优解更新完后，再更新全局历史最优解，随后继续进行迭代，得到新的速度和位置，当迭代一定次数之后便得到较优解。



• 求解步骤

1、初始化参数

该函数共有三个变量x_1, x_2, x_3，空间维度为3，设粒子数量为N，在k次迭代中，第i个粒子的位置为$x_i^k=\left(x_{i1}^k, x_{i2}^k, x_{i3}^k\right)$，速度为$v_i^k=\left(v_{i1}^k, v_{i2}^k, v_{i3}^k\right), i=1,2,\ldots,N$，其初始位置为随机的。

2、计算适应度

适应度即该模型的目标函数值，对于该函数，若粒子坐标为：(x_1, x_2, x_3)，则适应度为：

$$y=-0.0003392 x_1^2-0.0001212 x_2^2+0.1069 x_1+0.064 x_2+\frac{3.829 x_3-2082}{x_3+85.62}+73361$$

3、求初始全局最优解 第i个粒子的个体最优解记作pbest，通过比较全部粒子的pbest，找出最大值，记做全局最优解gbest。

4、更新粒子速度

$$v_i^k=w v_i^{k-1}+c_1\left(\mathrm{pbest}_i^{k-1}-x_i^{k-1}\right)+c_2\left(\mathrm{gbest}^{k-1}-x_i^{k-1}\right)$$

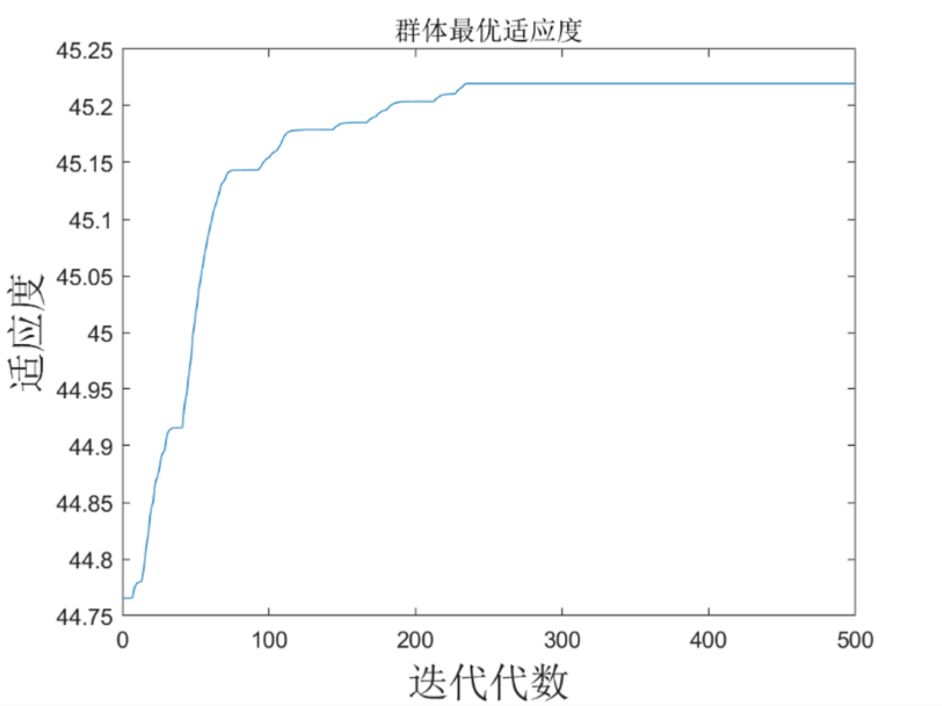
其中w, c_1, c_2分别表示惯性，个体经验和社会经验的权重，当$c_1 = 0$时，粒子是盲从的，不学习任何自身经验，当$c_2 = 0$时，粒子时自大的，不学习任何社会经验，当w偏大时，有利于全局搜索，函数收敛性差；当w偏小时，有利于局部搜索，但运算更慢，故理想状况下，前期w值应偏大，而到搜索后期，w值应逐渐减小。

5、更新位置（可行解）并求出适应度

6、更新个体最优解和全局最优解

7、迭代终止

• 结果分析



根据上述算法，求得当$x_1 = 290, x_2 = 260, x_3 = 699$时，该函数有最大值，故当每周理论培训时长为290分钟，实践培训时长为260分钟，投入699千元时，将有最多的参与人数，即有45.216%的人参加培训。

5.5 第三问的第三小问的求解

从应用价值上进行分析：结果的实用性：本论文分析所得结果能够直接应用于学校的教育、管理或服务流程中，提高教学效率和质量。

成本效益分析：考虑实施结果所需的投入与参赛人数之间的关系。可以优化资源配置，使得在合适的投入下能有较多的学生参赛。

可持续性：评估结果是否具有长期的应用潜力，以及是否能够在不断变化的环境中保持其价值。

可扩展性：考虑结果是否可以推广到学校的其他部门或者其他赛事。

改进措施： 数据准确性：检查数据收集和分析过程，确保结果的准确性和可靠性。

持续监控：建立机制以持续监控结果的应用效果，确保其持续符合学校的需求。

本论文经过计算所得的参赛人数取得最大值对应的三个变量值为：

$$x_1 = 290, x_2 = 260, x_3 = 699$$

每周理论培训时长为290分钟，实践培训时长为260分钟，投入699千元时，将有最多的参与人数。但是，考虑到实际情况，向学校相关部门提出以下更为合理的培训方案：

1、纸上得来终觉浅，绝知此事要躬行，建议适当增加实践教学时长，提高学生动手能力，建议把实践培训时长根据实际情况进行适当延长。

2、结合实际，考虑到学校经费等问题，且根据第一小问所计算出的函数曲线可以看出，当投入达到一定值时，贡献的增长率比起另外两个因素不算特别显著，所以建议适当削减学校投资经费。根据国家公务局《中央和国家机关培训费管理办法》[4]的综合定额标准，将投入设置在510千元左右较为合理。

六、 模型评价

6.1 模型优点

第一题模型采用秩和比评价方法进行模型建立，该方法适用性广泛，对指标的选择没有具体的需求，故适用于各类评价对象也包括该题；在计算过程中数据的秩次而非原始数据，这有助于消除异常值的影响，使得结果更为稳健；在排名时，依据数据大小的相对关系，而不是依赖于具体的数值大小，这使得评价更加客观，减少了量纲的影响。

第三题针对不同的问题建立不同的模型，如针对第一问，分别建立理论培训时间、实践培训时间、投入与参加人数的关系，而针对第二小问，则建立一个多元非线性回归函数进行问题分析，并考虑到现有求解非线性多元函数速度较慢，故提出用启发式算法针对该函数编程求解，使得其求解速度明显优化、加快收敛速度，增强全局搜索能力。

6.2 模型缺点

第二题的模型精度受样本量影响：在实际应用过程中，若选取的样本数量较少，则可能会导致熵值计算不准确，从而影响权重的计算和指标排序的结果。

七、 参考文献

[1] 田凤调. 秩和比法及其应用[J]. 中国医师杂志, 2002(2): 115-119.

[2] 司守奎, 孙玺菁. 数学建模算法与应用[M]. 3 版.北京: 国防工业大学出版社, 2023.

[3] 杨维, 李歧强. 粒子群优化算法综述[J]. 中国工程科学, 2004(5): 87-94.

[4] 财政部, 中共中央组织部, 国家公务员局. 国家公务员局《中央和国家机关培训费管理办法》[P/OL]. 2017-01-05. http://www.scs.gov.cn/zcfg/201701/t20170105_6894.html.

八、 附录

附录 1: B31.m 第一问求解每个因素与参加人数的关系

```
clc,clear,close all
% 理论培训时间与参加人数关系
x1 = table2array(readtable("C3.xlsx","Range","A2:A11"));
y1 = table2array(readtable("C3.xlsx","Range","D2:D11"));
scatter(x1,y1,'filled');
[b,bint,r,rint,stats]=regress(y1,[ones(10,1),x1,(x1.^2)])
rcoplot(r,rint);
x1(10,:)=[];
y1(10,:)=[];
[b,bint,r,rint,stats]=regress(y1,[ones(9,1),x1,(x1.^2)])
rcoplot(r,rint);
% 实践培训时间与参加人数关系
x2 = table2array(readtable("C3.xlsx","Range","B12:B21"));
y2 = table2array(readtable("C3.xlsx","Range","D12:D21"));
scatter(x2,y2,'filled');
y2(2)=(y2(1)+y2(3))/2;
scatter(x2,y2,'filled');
[b,bint,r,rint,stats]=regress(y2,[ones(10,1),x2,(x2.^2)])
rcoplot(r,rint);
y2(5)=(y2(4)+y2(6))/2;
y2(9)=(y2(8)+y2(10))/2;
[b,bint,r,rint,stats]=regress(y2,[ones(10,1),x2,(x2.^2)])
rcoplot(r,rint);
% 投入与参加人数关系
x3 = table2array(readtable("C3.xlsx","Range","C22:C31"));
y3 = table2array(readtable("C3.xlsx","Range","D22:D31"));
scatter(x3,y3,'filled');
hold on
ft = fittype( 'rat11' );
opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
opts.Display = 'Off';
opts.StartPoint = [0.237910506100943 0.558026701756326 0.299620025432728];
[fitresult, gof] = fit( x3, y3, ft, opts );
```

附录 2: B32.m 第二问求解总体回归方程模型

```
clc,clear,close all
data = table2array(readtable("C3.xlsx","Range","A2:D31"));  % 导入数据
std=y1(7);
% 理论培训时间与参加人数回归分析
x1 = data(1:10,1);
y1 = data(1:10,4);
for i = 1:10
    y1(i)=std-y1(i);
end
scatter(x1,y1,'filled');
hold on
ft = fittype( 'poly2' );
[fitresult, gof] = fit( x1, y1, ft );
plot( fitresult, x1, y1 );
xlabel( 'x1');
ylabel( 'y1');
legend(["x1",""])
grid on
hold off
% 实践培训时间与参加人数回归分析
x2 = data(11:20,2);
y2 = data(11:20,4);
for i = 1:10
```

```
    y2(i)=std-y2(i);
end
y2(2)=(y2(1)+y2(3))/2;
scatter(x2,y2,'filled');
hold on
r=polyfit(x2,y2,2);
t=0:1:400;
yy = r(1)*t.^2+r(2).*t+r(3);
plot(t,yy,'-');
grid on
hold off
% 投入与参加人数回归分析
x3 = data(21:30,3);
y3 = data(21:30,4);
for i = 1:10
    y3(i)=std-y3(i);
end
scatter(x3,y3,'filled');
hold on
ft = fittype( 'rat11' );
opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
opts.Display = 'Off';
opts.StartPoint = [0.350727103576883 0.939001561999887 0.875942811492984];
[fitresult, gof] = fit( x3, y3, ft, opts );
h = plot( fitresult, x3, y3 );
xlabel( 'x3');
ylabel( 'y3');
legend(["x3",""])
grid on
hold off
% 回归方程检验
y = data(:,4);
t=1:1:30;
hold on
y = data(1:30,4);
y(12)=(y(11)+y(13))/2;
sst=0; %SST
for i = 1:30
    sst=sst+(y(i)-sum(y)./30)^2;
end
ssr=0; %SSR
for i = 1:30
    ssr=ssr+(yf(x1(i),x2(i),x3(i))-sum(y)./30)^2;
end
sse=0; %SSE
for i = 1:30
    sse=sse+(y(i)-yf(x1(i),x2(i),x3(i))).^2;
end
rsquare=1-sse/sst   % 相关系数检验

msr=ssr/(3-1);
mse=sse/(30-3);
f=msr/mse   %f 检验
```

附录 3: yf.m 第二问回归函数

```
function t = yf(i,j,k)
    t=45.15-(0.0003392*i.^2-0.1969*i+28.4 ...
      +0.0001212184355522038*j.^2-0.064028817370481.*j+9.417872659555059+...
      (-3.829*k+2082)./(k+85.62));
end
```

附录 4:　PSO.m 粒子群算法求解最大值

```
clc,clear,close all

w=0.5;       % 惯性
c1 = 0.8;       % 个体经验
c2 = 0.9;       % 社会经验
maxg = 500;     % 迭代次数
N = 600;        % 粒子数
Vmax = 1;       % 速度范围
Vmin = -1;
Xmax = 800;     % 变量取值范围
Xmin = 0;
dim = 3;        % 函数表达式的自变量个数
for i=1:N
    location(i,:)=Xmax*abs(rands(1,dim));     % 初始化坐标
    V(i,:)=Vmax*rands(1,dim);                 % 初始化速度
    fitness(i)=yf(location(i,1),location(i,2),location(i,3));      % 适应度
end
[fitnessgbest,bestindex]=max(fitness);%fitnessgbest 是全局最优解对应的适应度
gbest=location(bestindex,:);    % 全局最优解
pbest=location;                 % 所有粒子的个体最优解
fitnesspbest=fitness;           % 所有粒子的个体最优解对应的适应度
for i=1:maxg
    for j=1:N
        % 根据惯性、个体最优 pbest 和群体最优 gbest 并更新速度
        V(j,:) = w*V(j,:) + c1*(pbest(j,:) - location(j,:)) + c2*(gbest - location(j,:));
        for k = 1:dim     % 限制速度不能过大
            if V(j,k) > Vmax
                V(j,k) = Vmax*0.9;
            elseif V(j,k) < Vmin
                V(j,k) = Vmin*0.9;
            end
        end
        location(j,:) = location(j,:) + V(j,:); % 更新位置
        for k = 1:dim       % 限制位置不能超过边界
            if location(j,k) > Xmax
                location(j,k) = Xmax*0.9;
            elseif location(j,k) < Xmin
                location(j,k) = Xmin*0.9;
            end
            if location(j,3) > 700
                location(j,3) = 650;
            end
        end
        %更新第 j 个粒子的适应度值
        fitness(j) = yf(location(j,1),location(j,2),location(j,3));
    end
    for j = 1:N
        if fitnesspbest(j) < fitness(j)     %更新个体最优解
            pbest(j,:) = location(j,:);
            fitnesspbest(j) = fitness(j);
        end
        if fitnessgbest < fitness(j)    %群体最优更新
            gbest = location(j,:);
            fitnessgbest = fitness(j);
        end
    end
    yy(i) = fitnessgbest;
end
figure;
plot(yy)
title('群体最优适应度','fontsize',12);
xlabel('迭代代数','fontsize',18);ylabel('适应度','fontsize',18);
```