所属类别	2023 年“华数杯”全国大学生数学建模竞赛	参赛编号
研究生组		CM2302421

母亲身心健康对婴儿成长的影响

摘要

母亲是婴儿成长过程中最为亲密的人，母亲的身心健康对婴儿的成长影响巨大，为婴儿提供身体上的养育和情感上的支持。本文针对母亲的身心健康对婴儿的行为特征和睡眠质量展开研究。

对于问题一，首先对问卷数据进行**数据预处理**，包括对异常值的筛除，进而将对指标进行分类，母亲的指标分为生理指标和心理指标两大类，婴儿的指标分为婴儿行为特征和睡眠质量。分别对以上指标进行**描述性统计**和可视化分析，大致判断母亲身心健康指标与婴儿行为特征、睡眠质量之间是否具有相关性，在此基础上运用**皮尔逊相关系数**检验母亲的身体指标变量和心理指标与婴儿的行为特征变量和睡眠质量变量之间的相关性，分析得到（1）**母亲年龄、CBTS、EPDS、HADS 与婴儿行为特征之间存在显著的相关关系（相关系数依次为-0.115、0.086、0.121、0.110）。（2）CBTS、EPDS、HADS 与婴儿整晚睡眠时间之间存在显著的负相关关系（相关系数依次为-0.115、-0.224、-0.151）。（3）妊娠时间、EPDS、HADS 与婴儿睡醒次数之间存在显著的正相关关系（相关系数依次为 0.084、0.141、0.116）。**

对于问题二，将预处理后的表格数据进行问卷的**可靠性检验**，利用 **SPSS** 检验了相关性较强的 CBTS、EPDS 与 HADS 三项心理指标，通过可靠性统计与 **KMO** 和**巴特利特检验**得出克隆巴赫值为 0.902，KMO 量为 0.728，这说明研究方法和结果具有可靠性。对婴儿的行为特征结果进行数据编码处理，进行定量化转换。根据母亲的身体指标与心理指标，与行为特征建立基于 **AdaBoost 算法**的婴儿行为特征分类模型，通过**交叉验证**与**混淆矩阵检验**，准确率为 64.1%，分类效果较好，最后利用训练完成的模型实现对最后有 20 组婴儿的行为特征信息的预测。

对于问题三，使用优化模型，根据题目已知信息建立约束条件，构建 **CBTS、EPDS、HADS 治疗费用最低**的目标函数，求得使婴儿行为特征转变为中等型和安静型的费用方案。

对于问题四，本文选择 **Critic 权重法**来为整晚睡眠时间、睡醒次数和入睡方式赋予权重。通过使用 **Spearman 相关性分析**，发现入睡方式与整晚睡眠时间呈正相关，与睡醒次数呈负相关，将入睡方式、整晚睡眠时间作为正向指标，睡醒次数作为负向指标，通过 **TOPSIS 优劣解距离法**得到睡眠指标综合评分，最后利用 **K-means** 方法对综合评分划分为优、良、中、差四类。利用 SPSS 的**偏最小二乘法**回归方法得到整晚睡眠时间、睡醒次数与入睡方式关于母亲身体指标和心理指标的回归方程，利用 **KNN** 对睡眠质量指标分类，求得 20 组婴儿的综合睡眠质量。

对于问题五，在第三问和第四问基础上，增加对婴儿睡眠质量的综合评级为优的约束条件，构建新的**目标函数**，建立新的**优化模型**进行求解治疗费用。

关键词：相关性分析；行为特征；多元线性回归模型；分类模型；优化模型

一、问题重述

1.1 研究背景

婴儿期是人一生中身心发展最迅速的时期，处于此时期的婴儿发展潜能巨大，其生理状况和心理发展易受外界影响，尤其是对于认知、情感和社会行为的可塑性强，因此，婴儿期是体格、智力和行为习惯养成的关键时期[1]。母亲作为婴儿成长过程中联系最为密切的人，承担了喂养儿童和启蒙教育儿童的重要责任，为婴儿提供情感支持和安全感。因此，母亲的身心健康对婴儿的身心发展具有重要影响，若母亲的心理健康状态处于不良状况，如抑郁、焦虑、压力等，可能会对婴儿的认知、情感和社会行为等方面产生负面影响。压力过大的母亲可能会对婴儿的生理和心理发展产生负面影响，例如影响其睡眠等方面。

以 390 名 1 至 3 个月的婴儿及其母亲为研究对象，结合母亲身体指标和心理指标的相关数据，其中，身体指标涵盖年龄、婚姻状况、教育程度、妊娠时间、分娩方式等诸多因素，心理指标包括产妇心理指标 CBTS（分娩相关创伤后应激障碍问卷）、EPDS（爱丁堡产后抑郁量表）、HADS（医院焦虑抑郁量表）。在此基础上研究其对婴儿身心成长的影响，如对婴儿睡眠质量指标包括整晚睡眠时间、睡醒次数和入睡方式的影响。

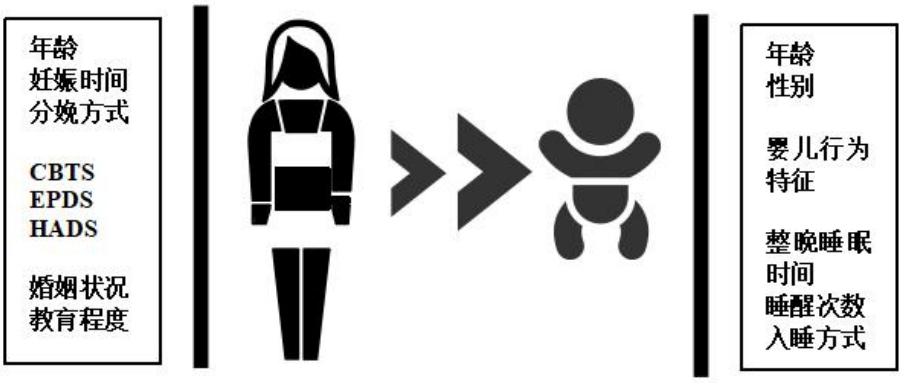


图 1 母亲身心健康指标与婴儿成长状况

1.2 问题提出

在查阅相关文献，了解专业背景的基础上，根据题目建立数学模型，回答下列问题。

1. 许多研究表明，母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量有影响，根据附件中的数据研究母亲的身心发展状况对婴儿的行为特征和睡眠质量之间是否存在规律，以及是何种规律。

2. 婴儿行为问卷是一个用于评估婴儿行为特征的量表，其中包含了一些关于婴儿情绪和反应的问题。我们将婴儿的行为特征分为三种类型：安静型、中等型、矛盾型。请你建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型。数据表中最后有 20 组（编号 391−410 号）婴儿的行为特征信息被删除，请你判断他们是属于什么类型。

3. 对母亲焦虑的干预有助于提高母亲的心理健康水平，改善母婴交互质量，促进婴儿的认知、情感和社交发展。CBTS、EPDS、HADS 的治疗费用相对于患病程度的变化率均与治疗费用呈正比，经调研，给出了两个分数对应的治疗费用（表 1）。现有一个行为特征为矛盾型的婴儿，编号为 238。请你建立模型，分析最少需要花费多少治疗费用，能够使婴儿的行为特征从矛盾型变为中等型？若要使其行为特征变为安静型，治疗方案需要如何调整？

表 1 病患得分与治疗费用

CBTS		EPDS		HADS	
得分	治疗费用（元）	得分	治疗费用（元）	得分	治疗费用（元）
0	200	0	500	0	300
3	2812	2	1890	5	12500

4. 婴儿的睡眠质量指标包含整晚睡眠时间、睡醒次数、入睡方式。请你对婴儿的睡眠质量进行优、良、中、差四分类综合评判，并建立婴儿综合睡眠质量与母亲的身体指标、心理指标的关联模型，预测最后 20 组（编号 391-410 号）婴儿的综合睡眠质量。

5. 在问题三的基础上，若需要让 238 号婴儿的睡眠质量评级为优，请问问题三的治疗策略是否需要调整？如何调整？

二、问题分析

结合题目背景，以 390 名 1-3 个月的婴儿及其母亲作为研究对象，研究母亲的身心健康指标和婴儿的成长状况指标，数据涵盖各种主题，如母亲的身体指标包括年龄、婚姻状况、教育程度、妊娠时间、分娩方式，以及产妇心理指标 CBTS、EPDS 和 HADS，婴儿身心发展指标包括婴儿行为特征、睡眠质量等，以此分析母亲的身心健康状况对婴儿成长的影响。

（1）母亲的相关数据指标：

- 身体指标：年龄、妊娠时间、分娩方式
- 心理指标：CBTS（分娩相关创伤后应激障碍问卷）、EPDS（爱丁堡产后抑郁量表）、HADS（医院焦虑抑郁量表）。
- 其他指标：婚姻状况、教育程度

（2）婴儿的相关数据指标：

- 年龄
- 性别
- 婴儿行为特征
- 睡眠质量：整晚睡眠时间、睡醒次数、入睡方式

在对问题进行整体解读与分析后，明确了问题的大方向为探究母亲身心健康对婴儿成长的影响，在此基础上，逐一对每一小问进行详细分析，分析问题实质，梳理解题思路，最终找出求解问题的的最佳算法和模型，得到较为可信的结果。

2.1 问题一的分析

针对问题一，要求研究母亲的身体指标和心理指标对婴儿的行为特征和睡眠质量的影响。第一步，对所有变量进行描述性统计，得到整体样本的分布情况。第二步，分为母亲的身体指标和心理指标对婴儿的行为特征的影响与母亲的行为指标和心理指标对婴儿的睡眠质量的影响两个部分进行讨论，用可视化工具描述在不同婴儿行为特征下母亲的身体和心理指标情况，以及不同睡眠质量变量条件下母亲的身体和心理指标情况，观察其是否有影响。第三步，运用皮尔逊相关系数检验母亲的身体指标变量和心理指标与婴儿的行为特征变量和睡眠质量变量之间的相关性，找到存在显著相关关系的变量。

2.2 问题二的分析

针对问题二，该题属于典型的分类问题，根据母亲的身心健康指标探究婴儿的行为

特征，对其进行分类。首先，在第一问的基础上，确定对婴儿行为特征影响显著的母亲的生理指标和心理指标，再结合筛选出来的显著性指标所对应的数据进行问卷的信度和效度检验，若问卷所导出的数据通过了信度和效度检验，则说明该研究方法和研究结果具有一定的有效性和准确性。由于婴儿的行为特征：安静型、中等型和矛盾型属于未定类数据，需对该指标进行数据编码，从而实现对婴儿行为特征的定量研究。最后，采用多分类模型建立婴儿行为特征与母亲的身体指标和心理指标的关系模型，若模型准确率较高，则在此基础上，带入最后 20 组（编号 391-410 号）母亲的相应身心健康指标预测其所对应的婴儿的行为特征。

2.3 问题三的分析

问题三为典型的优化求解问题，在第二问在明确母亲心理指标与婴儿行为特征关系的基础上，通过建立多元线性回归模型得到婴儿行为特征与母亲年龄、CBTS、EPDS 和 HADS 的函数表达式，再根据 CBTS、EPDS、HADS 的治疗费用相对于患病程度的变化率均与治疗费用呈正比这一规律，将治疗费用设定为以上三种心理指标的加权和，权重为各自变化率。通过题目要求设定约束条件，根据三项心理指标和治疗费用构造目标函数，建立优化模型实现治疗费用最小化将婴儿行为特征由矛盾型转化为中等型和安静型。

2.4 问题四的分析

问题四属于综合评价和分类问题，在对婴儿的睡眠质量进行综合评价之前，选择 Critic 权重法来为整晚睡眠时间、睡醒次数和入睡方式赋予权重。通过使用 Spearman 相关性分析，得到入睡方式与整晚睡眠时间和睡醒次数之间的相关关系，通过 TOPSIS 优劣解距离法得到睡眠指标综合评分，最后利用 K-means 方法对综合评分划分为优、良、中、差四类。

在此基础上，利用 SPSS 的偏最小二乘法回归方法得到整晚睡眠时间、睡醒次数与入睡方式关于母亲身体指标和心理指标的回归方程，将最后 20 组母亲的身体指标与心理指标数据代入回归方程，求解得到婴儿睡眠质量指标，最后利用 KNN 对睡眠质量指标分类，得到最后 20 组婴儿的综合睡眠质量。

2.5 问题五的分析

第五问关注的是婴儿睡眠质量与母亲的三项心理指标之间的相关关系，在确定婴儿睡眠质量与 CBTS、EPDS 和 HADS 的相关关系的基础上，利用第三问中求解出的三项心理指标的治疗费用的函数关系式，在问题三的基础上增加一项使婴儿睡眠质量为优的约束条件，使用优化算法，构建目标函数，利用 Lingo 进行求解，并对结果做出解释，得出使 238 号婴儿的综合睡眠质量为优时的治疗策略。

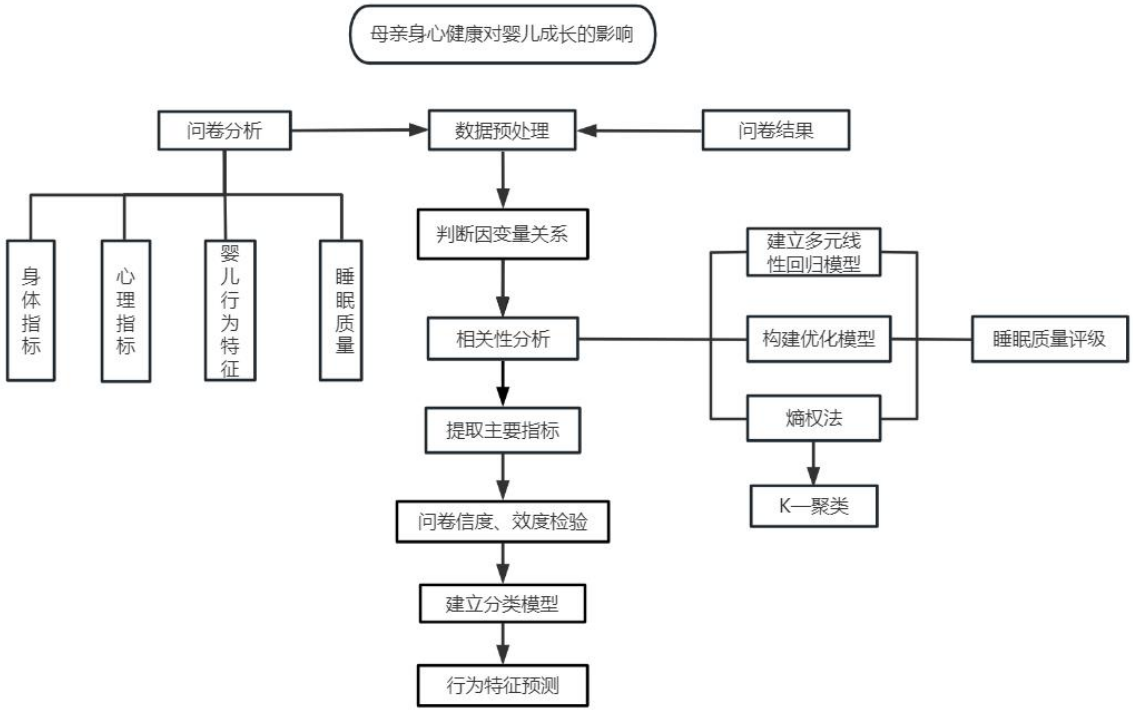


图 2 问题分析流程图

三、条件假设

为实现模型建立和求解的可靠性，提高结果的准确性，围绕本题假设以下条件：

1. 假设问卷数据在剔除异常值等数据预处理的操作后符合实际情况，可带入模型和算法中求得结果。
2. 假设对婴儿行为特征和睡眠质量的编码符合逻辑，对其分类科学合理。
3. 假设在进行综合评价时，对综合睡眠质量
4. 假设婴儿未使用影响睡眠质量的相关药物。
5. 假设婴儿的睡醒方式与睡醒次数呈负相关，与整晚睡眠时间呈正相关。
6. 假设婴儿被调查时未发现其他躯体疾病或精神问题。

四、符号说明

为方便模型建立和求解，对以下关键符号进行说明。

符号	说明
εt	第 t 轮迭代中分类器
ht	在训练数据集上的错误率
alpha_t	第 t 轮迭代中分类器 ht 的权重
H_t(x)	第 t 轮迭代中分类器
ht	对输入样本 x 的预测结果

x_1	CBTS 的治疗值
x_2	EPDS 的治疗值
x_3	HADS 的治疗值
y_1	CBTS 的治疗费用
y_2	EPDS 的治疗费用
y_3	HADS 的治疗费用
w_1	婴儿行为特征为中等型的治疗费用
w_2	婴儿行为特征为安静型的治疗费用

注：部分符号在首次引入时的模型建立与求解部分作出说明。

五、模型的建立与求解

5.1 问题一的统计与求解

5.1.1 数据预处理

在进行问题求解之前，有必要对数据进行有效性排查，进行数据预处理，包括对于数据的缺失值进行查找、补充，剔除异常值，对未定类的数据进行编码等。利用 Excel 中的 COUNTA 函数查找问卷结果中的空缺值，结果为原始数据不存在空缺值，只需要手动检查剔除异常值即可。

对于婴儿整晚睡眠时间，经筛查发现，编号为 180 的样本的婴儿整晚睡眠时间为 99:99 小时，严重脱离实际情况，因此认定样本编号为 180 的整晚睡眠时间为异常值，应被剔除。

对于母亲的婚姻状况，问卷规定了 1 为未婚，2 为已婚，因此问卷结果只能为其中之一，对不符合这一结果的样本数据进行筛选，作为异常值剔除，如表 5.1 所示。

表 5.1 婚姻状况异常值统计表

编号	母亲年龄	婚姻状况
43	35	3
95	33	3
134	35	3
196	34	3
231	31	6
301	26	3
306	26	6
308	38	3
355	31	3

5.1.2 描述性统计

在进行数据预处理的基础上，针对样本数据首先进行混合样本变量分布的描述性统计，其次按婴儿的行为特征分为安静型、中等型、矛盾型进行描述统计，以对比安静型、中等型、矛盾型的婴儿行为特征异质性差异。

（1）样本基本状况

表 5.1 展示了全样本母亲的身体指标、心理指标、婴儿行为特征和睡眠质量指标的描述性统计。由表 5.1 可知，在全样本中母亲年龄平均为 30.115 岁，妊娠时间平均为 39.146 周，正常妊娠时限为 40 周，妊娠时间状况良好。在分娩方式方面，选择自然分娩的有 394 人，占据整个样本的 98.75%，选择刨宫产的比例仅为 1.25%，自然分娩仍然是大部分女性普遍选择的生产方式。在婚姻状况方面，已婚母亲有 385 人，占据整个样本的 96.49%，未婚母亲比例较低，为 3.51%。从从受教育程度来看，大部分母亲的受教育程度较高，有 31.58%的母亲学历为高中及以下，大部分母亲学历在大学本科以上，其中研究生学历的比例最大，为 46.87%。

在母亲的心理指标方面，三种心理指标 CBTS（分娩相关创伤后应激障碍问卷）、EPDS（爱丁堡产后抑郁量表）、HADS（医院焦虑抑郁量表）得分均值分别为 5.972、9.087 和 7.899，标准差较大，分别为 4.981、6.724 和 4.276，说明不同母亲之间心理状况差别较大。

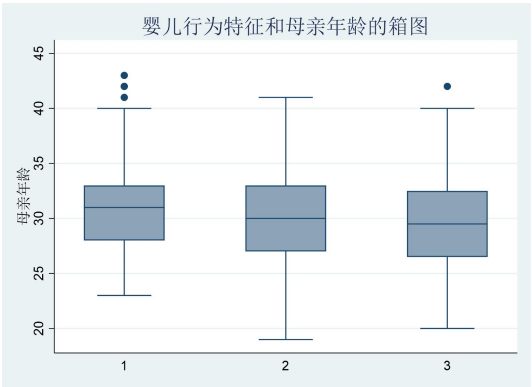
在婴儿的行为特征方面，中等型的婴儿有 219 人，占比为 57.78，安静型的婴儿有 116 人，占比为 30.61%，矛盾型婴儿比例最少，为 11.61%。

在婴儿的睡眠质量指标方面，婴儿整晚睡眠时间平均为 10.167 小时，婴儿睡醒次数平均为 1.424 次，睡眠状态总体较好。在入睡方式方面，五种入睡方式中，通过调整婴儿睡眠环境的环境营造法所占比例最大，为 43.01%，哄睡法、抚触法、安抚奶嘴法、定时法所占比例分别为 21.90%、17.94%、5.28%、11.87%。

表 5.2 样本基本情况表

	变量	均值	标准差	最小值	最大值
母亲的身体指标	母亲年龄	30.115	4.292	19	43
	妊娠时间周数	39.146	1.881	26.5	43
	分娩方式	1.012	0.111	1	2
	婚姻状况	1.964	0.184	1	2
	教育程度	4.080	1.004	1	5
母亲的心理指标	CBTS	5.972	4.981	0	21
	EPDS	9.087	6.724	0	28
	HADS	7.877	4.276	0	20
	婴儿行为特征	1.810	0.622	1	3
婴儿睡眠质量指标	整晚睡眠时间	10.167	1.447	5	12
	睡醒次数	1.472	1.623	0	10
	入睡方式	3.050	1.401	1	5

（2）婴儿行为特征的母亲身体和心理指标情况



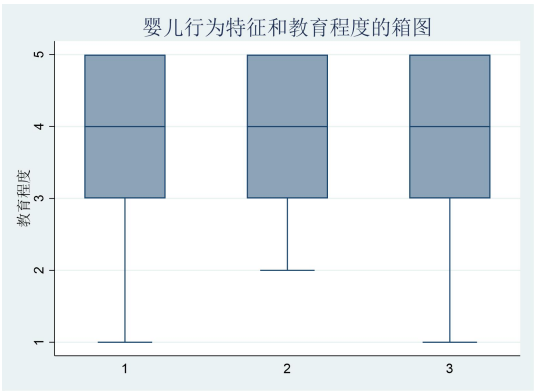
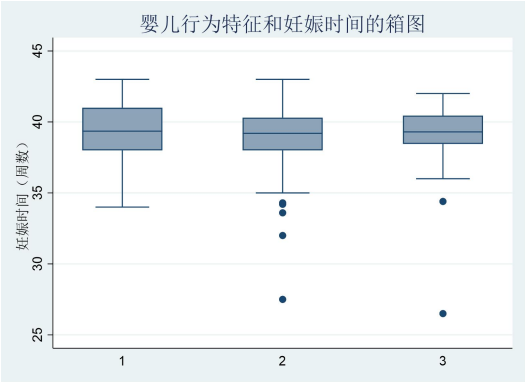


图 1 婴儿行为特征和母亲年龄的箱图　　**图 2 婴儿行为特征和教育程度的箱图**

由图 1 可知，不同婴儿行为特征下母亲年龄的中位数不同，其中婴儿行为特征为安静型的母亲年龄的中位数最高，其次是中等型，最后是矛盾型，认为母亲年龄对婴儿行为特征有影响。由图 2 可知，教育程度在三种婴儿行为特征下的中位数相同，且箱体发小相同，认为母亲教育程度对婴儿行为特征无影响。



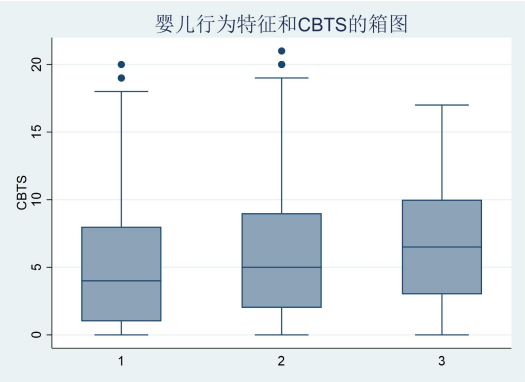
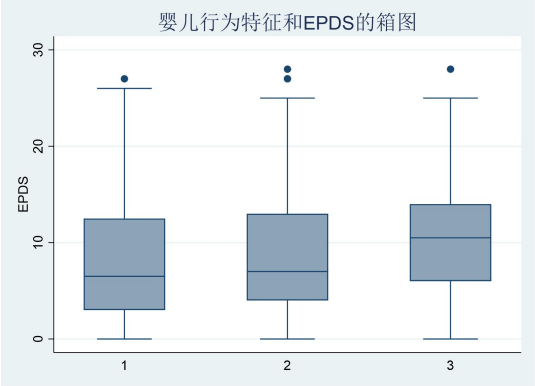


图 3 婴儿行为特征和妊娠时间的箱图　　**图 4 婴儿行为特征和 CBTS 的箱图**

由图 3 可知，妊娠时间在三种婴儿行为特征下的中位数几乎相同，认为妊娠时间对婴儿行为特征无影响。由图 4、图 5、图 6 可以看到 CBTS、EPDS、HADS 的中位数在三种婴儿行为特征下均不同，且矛盾型婴儿在三种母亲的心理指标分数的中位数最高，其次是中等型婴儿，最后是安静型婴儿。心理指标分数越高，女性的心理状况焦虑程度越高，婴儿更加倾向于矛盾型。因此，认为母亲的心理指标对婴儿行为特征有影响。



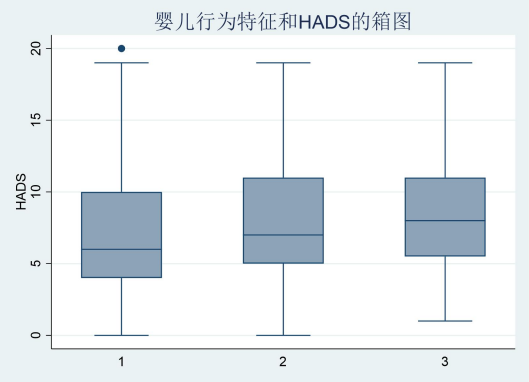
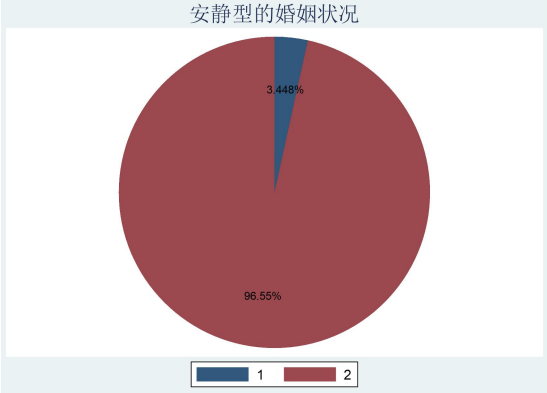


图 5 婴儿行为特征和 EPDS 的箱图　　**图 6 婴儿行为特征和 HADS 的箱图**



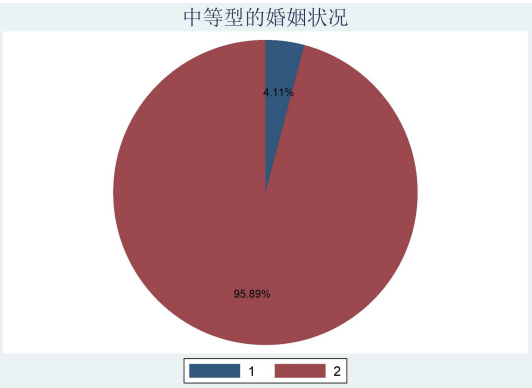


图 7 安静型婴儿母亲婚姻状况的饼图　　**图 8 中等型婴儿母亲婚姻状况的饼图**

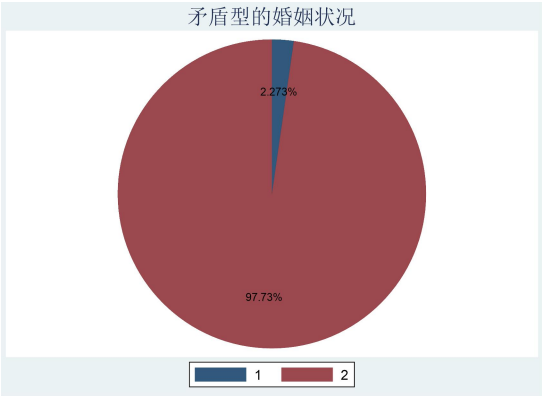


图 9 矛盾型婴儿母亲婚姻状况的饼图

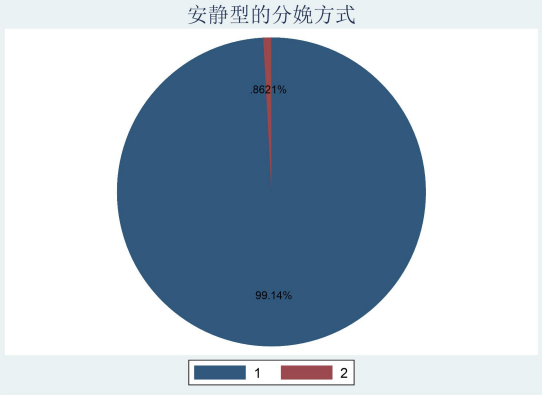


图 10 安静型婴儿母亲分娩方式的饼图

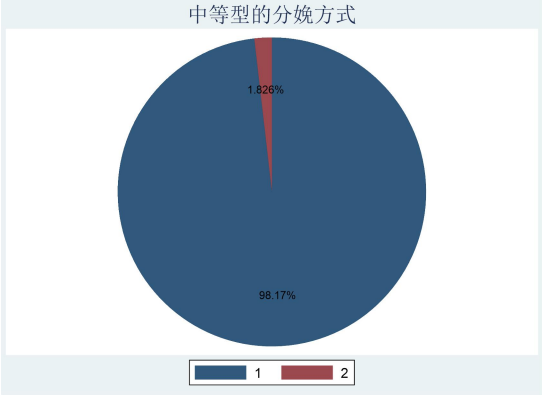


图 11 中等型婴儿母亲分娩方式的饼图

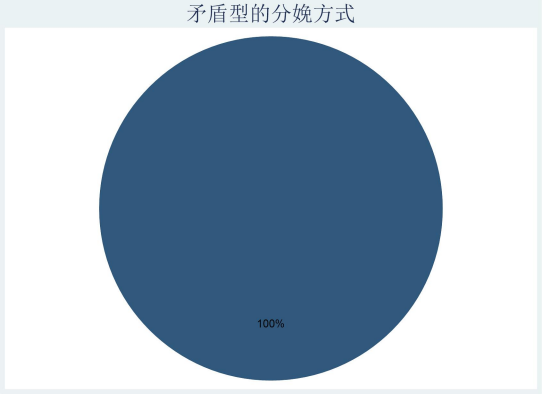


图 12 矛盾型婴儿母亲分娩方式的饼图

由图 7、图 8、图 9 可知，安静型、中等型、矛盾型婴儿的母亲婚姻状况中已婚和未婚比例相差不大，认为母亲婚姻状况对婴儿行为特征无影响。同样由图 10、图 11、图 12 可以观察到安静型、中等型、矛盾型婴儿的母亲分娩方式中自然分娩和刨宫产比例相差不大，认为母亲分娩方式对婴儿行为特征无影响。

由以上箱型图和饼图的可视化可以初步判断，母亲婚姻状况和母亲的心理指标对婴儿行为特征有影响。

5.1.2 相关性分析

通过对变量之间进行相关系数检验，发现母亲年龄、CBTS、EPDS、HADS 与婴儿行为特征之间存在显著的相关关系（相关系数依次为-0.115、0.086、0.121、0.110）。心理问卷得分越高，表明女性的心理状况越焦虑，婴儿更倾向于矛盾型，反之则倾向于安静型。CBTS、EPDS、HADS 与婴儿整晚睡眠时间之间存在显著的负相关关系（相关系数依次为-0.115、-0.224、-0.151）。女性的心理状况越焦虑，婴儿整晚睡眠时间越短。妊娠时间、EPDS、HADS 与婴儿睡醒次数之间存在显著的正相关关系（相关系数依次为 0.084、0.141、0.116）。母亲年龄与入睡方式之间在 10%的统计水平下存在负相关关系（相关系数为-0.097）。

故得到结论，母亲年龄、母亲的心理指标对婴儿行为特征有影响；母亲年龄、妊娠时间、心理指标对婴儿睡眠质量有影响。

表 5.3 相关系数表

变量	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
母亲年龄	1.000											
婚姻状况	0.027	1.000										
教育程度	0.228***	0.124**	1.000									
妊娠时间	-0.054	-0.013	0.021	1.000								
分娩方式	0.039	0.021	-0.009	-0.221***	1.000							
CBTS	-0.050	-0.034	-0.111**	-0.123**	-0.031	1.000						
EPDS	-0.125**	-0.034	-0.128**	-0.049	-0.028	0.784***	1.000					
HADS	-0.088*	-0.072	-0.122**	-0.097*	-0.065	0.686***	0.791***	1.000				
婴儿行为特征	-0.115**	0.008	0.040	-0.026	-0.002	0.086*	0.121**	0.110**	1.000			
整晚睡眠时间	0.041	0.008	-0.009	0.057	0.019	-0.115**	-0.224***	-0.151***	-0.097*	1.000		
睡醒次数	0.033	-0.038	0.071	0.084*	-0.062	0.059	0.141***	0.116**	0.259***	-0.293***	1.000	
入睡方式	-0.097*	-0.053	-0.036	0.013	0.062	0.050	0.006	0.055	-0.019	0.272***	-0.269***	1.000

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

5.2 问题二的统计与求解

5.2.1 问卷可靠性检验

问题一通过对各变量进行相关性分析，研究了母亲身体指标和心理指标一系列指标对婴儿的行为特征和睡眠质量的影响方向和程度，筛选出相关性强的一系列指标：CBTS（分娩相关创伤后应激障碍问卷）、EPDS（爱丁堡产后抑郁量表）和 HADS（医院焦虑抑郁量表），在此基础上，利用 SPSS 检验 CBTS、EPDS 和 HADS 这三项指标的信度和效度，判断问卷结果是否具有可靠性。通过可靠性统计得到克隆巴赫值为 0.885，基于标准化项的克隆巴赫值为 0.902，说明问卷具有很高的可信度；对问卷进行 KMO 和巴特利特检验，得到 KMO 量为 0.728，符合 KMO 量＞0.6 且显著性＜0.05，问卷通过效度检验，该研究方法和结果具有准确性。

表 5.4 可靠性检验表

克隆巴赫 Alpha	基于标准化项的克隆巴赫 Alpha	项数	KMO 取样适切性量数	巴特利特球形度检验		
.885	.902	3	.728	近似卡方	自由度	显著性
				778.479	3	.000

5.2.2 数据编码处理

婴儿的行为特征和整晚睡眠时间为未定类数据，需对其进行编码处理，编码原则为 1，2，3 分别代表婴儿行为特征中的安静型、中等型和矛盾型，在此基础上建立数学模型，对婴儿的行为特征进行定量研究。对整晚睡眠时间进行数据编码，以半个小时为 0.5 个数值，整晚睡眠时间对应的编码数据依次为 5，6，7，7.5，8，8.5，9，9.5，10，10.5，11，11.5，12，11.25。通过对数据进行编码处理，将两大指标数据转化为定类数据，为建立相应数学模型，进一步探究母亲身心健康指标对婴儿成长的影响具有基础性作用。

5.2.3 基于 AdaBoost 集成学习算法的三分类模型的建立

通过对婴儿行为特征进行标签编码后，确定其为定类变量。在建立婴儿的行为特征与母亲的身体指标与心理指标的关系模型时，选择身体指标（母亲年龄、婚姻状况、教育程度、妊娠时间、分娩方式）与心理指标（CBTS、EPDS、HADS）作为特征矩阵，婴儿行为特征作为目标类别，将样本特征筛选并将数据预处理为 X，y 的数据集，其中数据集 X 为包含各项身体指标与心理指标的特征矩阵，数据集 y 为婴儿行为特征对应的类别标签。

在本题中，婴儿行为特征对应三种类别的分类可以采用分类模型解决，根据对多种分类模型的实验，最终发现 AdaBoost 算法通过组合多个弱学习器来形成一个强大的分类器，在小样本量、多分类问题中具有较高的分类准确率，效果较好。

（1）算法步骤

利用预处理后的数据集提取关键特征，构建母亲身体指标与心理指标的特征矩阵 X：

$$X = \{(x_1, x_2, x_3, ...), (m_1, m_2, m_3, ...), ..., (n_1, n_2, n_3, ...)\}$$

构建婴儿行为特征的类别矩阵 y：

$$y = \{(y_1, y_2, y_3, ...)\}$$

Step1 数据准备：其中x_i,m_i,...,n_i分别表示在不同特征下的第 i 个样本的特征向量，y_i是第 i 个样本对应的标签类别，该问题中婴儿行为特征有三个类别，将问题转换为三个二分类问题，即每次将安静型、中等型、矛盾型中一个类别作为正例，其余两个类别作为负例。

Step2 初始化样本权重：对于上述三个二分类任务中的每一个，记初始化样本权重w_i为：

$$w_i = \frac{1}{n}$$

其中 n 是样本数量，在初始化权重阶段，每个样本的权重都相等。

Step3 迭代训练：对于 t = 1 到 T（T 是迭代次数），对于每个二分类任务执行以下步骤：

a. 使用当前样本权重w_t训练一个弱分类器h_t：训练的弱分类器是简单的决策树或逻辑回归等，该弱分类器是一个比较简单的分类器。

b. 计算分类器h_t在训练集上的错误率ε_t：计算分类器h_t在训练数据集上的加权错误率，其中样本权重w_t会影响每个样本的重要性。计算得到的错误率ε_t表示为：

$$\varepsilon_t = \frac{\Sigma(w_t \cdot I(y_i \neq h_t(x_i)))}{\Sigma(w_t(i))}$$

其中 I(expression)是指示函数，当$y_i \neq h_t(x_i)$时为 1，否则为 0。

c. 计算分类器h_t的权重$alpha_t$：利用弱分类器的错误率ε_t计算分类器权重，记分类器权重$alpha_t$为：

$$alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

d. 更新样本权重：对于所有样本，更新样本权重w_{t+1}(i)为：

$$w_{t+1}(i) = w_t(i) \cdot \exp(-alpha_t \cdot y_i \cdot h_t(x_i))$$

这里，y_i是样本 i 的真实标签，$h_t(x_i)$是分类器h_t对样本 i 的预测结果。

e. 样本权重归一化：将样本权重w_{t+1}(i)规范化，使它们的总和为 1，以便它们仍然构成一个概率分布。

Step4 最终分类器的组合：在每个二分类任务的迭代完成后，将所有弱分类器的权重$alpha_t$和预测结果$h_t(x)$ 组合成最终的分类器 H(x)，记 H(x)的数学表达式为：

$$H(x) = \text{argmax}\{\Sigma rglpha\_t \cdot I(h_t(x) = c)\} \quad c \in \{1,2,3\}$$

对于输入样本 x，对每个二分类任务执行H_t(x)=sign(Σ(alpha_t*h_t(x)))，其中 sign()是符号函数，返回正数时为 1，负数时为-1。

根据H_t(x)的投票结果，选择获得最多正例票数的类别作为最终的分类标签。

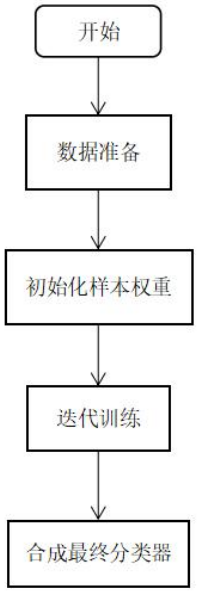


图 13 AdaBoost 算法流程图

（2）模型求解

根据表格中提供的数据，选择了母亲年龄、婚姻状况、教育程度、妊娠时间、分娩方式，以及 CBTS、EPDS、HADS 共八个特征参与特征矩阵的构建，选择婴儿行为特征作为特征矩阵对应的标签矩阵，利用 Python 建立基于 AdaBoost 算法的多分类模型来进行拟合预测，并通过循环进行多次模型训练。

最终发现，在数据划分过程中训练集与测试集比例在 0.7，随机种子设置为 2225，在创建 AdaBoost 分类器过程中，设置决策树最大深度为 4，弱分类器个数设置为 67，学习率设置为 1.0，随机种子设置为 42，模型准确率最高，为 64.1%。

将表格中最后 20 组（编号 391-410 号）母亲身体指标与心理指标作为特征矩阵代入到训练完成的模型中，分类预测得到婴儿的行为特征，如下表。

表 5.4 婴儿行为特征预测表

编号	391	392	393	394	395	396	397	398	399	400

预测值	2	2	2	1	2	2	2	2	2	2
所属类别	中等型	中等型	中等型	安静型	中等型	中等型	中等型	中等型	中等型	中等型
编号	**401**	**402**	**403**	**404**	**405**	**406**	**407**	**408**	**409**	**410**
预测值	2	2	2	1	2	1	2	2	2	2
所属类别	中等型	中等型	中等型	安静型	中等型	安静型	中等型	中等型	中等型	中等型

5.3 问题三的统计与求解

5.3.1 多元线性回归模型的建立

通过第一问的分析，婴儿的行为特征与母亲年龄、CBTS、EPDS 和 HADS 有相关性，利用 Stata 使用最小二乘法求解婴儿特征与母亲身心健康指标的函数关系，建立婴儿的特征行为与年龄、心理指标的多元回归模型。

（1）算法步骤

在问题一的基础上，利用最小二乘法对预处理后的母亲年龄、母亲心理指标：CBTS、EPDS 和 HADS 数据及其之间的相关关系构建多元线性回归模型，建立相关矩阵方程，基本思路如下：

设(x,y)是一对观测量，$x=[x_1,x_2,...,x_n]^T\in R_n$，$y=R$且满足以下理论函数：

$$y=f(x,w),$$

其中，$w=[w_1,w_2,...w_n]^T$为待定参数。

为寻找函数$f(x,w)$的参数 w 的最优估计值，对于给定 m 组（一般 m＞n）观测数据$(x_i,y_i)(i=1,2,...,m)$，求解目标函数$L\big(y,f(x,w)\big)=\sum_{i=1}^{m}[y_i-f(x_i,w_i)]^2$取最小值的参数$w_i(i=1,2,...,n)$。

（2）模型求解

将母亲的年龄、CBTS、EPDS 和 HADS 带入关系式中，将婴儿的行为习惯作 y，利用 Stata 计算回归系数，结果如下：

表 5.5 线性回归分析结果

变量	VIF	回归系数
母亲年龄	1.0200	-0.0145*
		(0.00741)
CBTS	2.6900	-0.00237
		(0.0103)
EPDS	3.8100	0.00805
		(0.00911)
HADS	2.7800	0.00646
		(0.0123)

常数项	2.136***
	(0.238)
样本数	379
R^2	0.026
P	0.0460

$^{*}p<0.1, ^{**}p<0.05, ^{***}p<0.01$

P 值为 0.046，经 F 检验，在置信水平为 95%的水平上具有显著性，拒绝回归系数为 0 的原假设，该多元回归模型基本符合要求。对于多重共线性问题，因 VIF＜10，不存在多重共线性问题，因此多元线性回归模型成立，根据得出的各变量的回归系数构建函数关系式：

$$y = 2.136 - 0.014x_{母亲年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS}$$

5.3.2 优化模型的建立

在建立母亲身心健康指标和婴儿成长关系的多元线性回归方程的基础上，以治疗费用最小为目标建立目标函数，以婴儿行为特征由矛盾型转变为中等型和安静型的阈值来设定约束条件，将治疗费用设定为 CBTS、EPDS 和 HADS 三个指标的加权和，权重为各自的变化率，在确定各项指标权重的基础上构造线性规划等优化模型求解。

CBTS、EPDS 和 HADS 的治疗费用与患病程度为正比关系，设 CBTS 的治疗值为x_1，EPDS 的治疗值为x_2，HADS 的治疗值为x_3，各自对应的治疗费用依次为y_1，y_2和y_3，治疗总费用为w，建立目标函数，使得 w 值最小。

建立有关治疗费用的函数关系式分别为：

$$y_1 = k_1x_1 + b_1$$
$$y_2 = x_2 + b_2$$
$$y_3 = k_3x_3 + b_3$$

建立治疗费用w的目标函数：

$$w = y_1 + y_2 + y_3$$

根据题目已知条件可以求出 CBTS、EPDS 和 HADS 的患病程度变化率分别为k_1 =870.67，k_2 =690，k_3 =2440，b_1 =200，b_2 =500，b_3 =300，可求得y_1，y_2，y_3的具体表达式：

$$y_1 = 870.67x_1 + 200$$
$$y_2 = 690x_2 + 500$$
$$y_3 = 2440x_3 + 300$$

在问题二中已经对婴儿的行为特征进行编码，安静型编码为 1，中等型为 2，矛盾型为 3，因此要想使矛盾型转化为中等型，则$y \le 2$，结合题目已知的原始数据，第 238 组母亲的 CBTS、EPDS 和 HADS 得分分别为 15、22 和 18，在此基础上，根据婴儿的行为特征的函数关系式与治疗费用构成目标函数如下：

$$y = 2.136 - 0.014x_{母亲年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS}$$

$$s.t.\begin{cases} y_1 = 870.67x_1 + 200 \\ y_2 = 690x_2 + 500 \\ y_3 = 2440x_3 + 300 \\ y = 2.136 - 0.014x_{年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS} \\ y \le 2 \\ x_{CBTS} = 15 - x_1 \\ x_{EPDS} = 22 - x_2 \\ x_{HADS} = 18 - x_3 \\ x_1, x_2, x_3 为整数 \end{cases}$$

利用 Lingo 得出结果：w_1=5380，x_2=7，即令第 238 组的婴儿行为特征由矛盾型转化为中等型的 CBTS、EPDS 和 HADS 治疗费用最小为 5830 元，才能实现矛盾性的婴儿行为特征转化为中等型。

同理，将矛盾型转化为安静型，设治疗费用为w_2，仅需改变约束条件$y \le 1$，根据已知条件建立目标函数如下：

$$s.t.\begin{cases} y_1 = 870.67x_1 + 200 \\ y_2 = 690x_2 + 500 \\ y_3 = 2440x_3 + 300 \\ y = 2.136 - 0.014x_{年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS} \\ y \le 1 \\ x_{CBTS} = 15 - x_1 \\ x_{EPDS} = 22 - x_2 \\ x_{HADS} = 18 - x_3 \\ x_1, x_2, x_3 为整数 \end{cases}$$

利用 Lingo 得出结果：w_2=92080，即治疗费用最低为 9208 元，才能实现将婴儿行为特征由矛盾型转化为安静型。

5.4 问题四的统计与求解

5.4.1 模型的建立

（1） Critic 赋权法

运用 Critic 赋权法综合影响因素的重要性和影响因素提供的信息量来确定影响各因素的权重，进而得到模型对相关因素的依赖性。Critic 法是基于评价指标的对比强度和指标之间的冲突性来综合衡量指标的客观权重。考虑指标变异大小的同时兼顾指标之间的相关性，完全利用数据自身的客观属性进行科学评价。

根据题干要求，现有婴儿睡眠质量，分别受整晚睡眠时间、睡醒次数和入睡方式的影响，

根据数据样本个数有 399 个，指标有 3 个，按照下面四个步骤分别计算三个睡眠质量指标的权重：

第一步，对每一个指标按照每个选项的数量进行归一化处理：
对于正向指标：

$$Y_{ij=}\frac{y_{ij}-\min(y_{1j},y_{2j}\cdots,y_{nj})}{\max(y_{1j},y_{2j}\cdots,y_{nj})-\min(y_{1j},y_{2j}\cdots,y_{nj})}\ (i=1,2\cdots399;\ j=1,2,3)$$

对于负向指标：

$$Y_{ij=}\frac{\max(y_{1j},y_{2j}\cdots,y_{nj})-y_{ij}}{\max(y_{1j},y_{2j}\cdots,y_{nj})-\min(y_{1j},y_{2j}\cdots,y_{nj})}\ (i=1,2\cdots399;\ j=1,2,3)$$

第二步，指标变异性，以标准差的形式来表现，计算各指标的标准差S_j：

$$\overline{Y_j}=\frac{\sum_{i=1}^{n}Y_{ij}}{n}$$

$$S_j=\sqrt{\frac{\sum_{i=1}^{n}(Y_{ij}-\overline{Y_j})^2}{n-1}}\qquad(n=399)$$

标准差越大表示该指标的数值差异越大，越能放映出更多的信息，该指标的评价强度也就越强，应该给该指标分配更多的权重。

第三步，指标冲突性，以相关系数的形式来表现，计算各指标之间的相关系数r_{ij}：

$$r_{ij}=\frac{\sum_{i=1}^{n}(y_i-\overline{y}_i)(y_j-\overline{y}_j)}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2\sum_{j=1}^{n}(y_j-\overline{y}_j)^2}\qquad(i=1,2,3;\ j=1,2,3)$$

其中，$\overline{y}_i$是指标y_i的平均指标值；$\overline{y}_j$是指标y_j的平均指标值。

$$R_j=n-\sum_{i=1}^{n}r_{ij}\qquad(n=399)$$

使用相关系数来表示指标间的相关性，与其他指标的相关性越强，则该指标就与其他指标的冲突性越小，反映出相同的信息越多，所能体现的评价内容就越有重复之处，一定程度上削弱了该指标的评价强度，应该减少对该指标分配的权重。

第四步，计算各指标所蕴含信息量C_j：

$$C_j=S_j(n-\sum_{i=1}^{n}r_{ij})=S_j\times R_j$$

第五步，得到各指标的权重W_j

$$W_j = \frac{C_j}{\sum_{j=1}^{n} C_j}$$

（2）TOPSIS 优劣解距离法

TOPSIS 优劣解距离法是一种综合评价方法，尽可能地保留了原始数据所包含的信息，进而较为精确地反映各个方案之间的差距。应将数据同向化和标准化，然后按照如下步骤：

第一步，敲定最优、最劣方案。最优方案Z^+由Z中每列元素的最大值构成，最劣方案Z^-由Z中每列元素的最小值构成，具体表达式如下：

$$\begin{aligned} Z^+ &= (\max\{z_{11}, z_{21} \cdots z_{n1}\}), \max\{z_{12}, z_{22} \cdots z_{n2}\}), \cdots, \max\{z_{1m}, z_{2m} \cdots z_{nm}\}) \\ &= (z_1^+, z_2^+ \cdots z_m^+) \end{aligned}$$

$$\begin{aligned} Z^- &= (min\{z_{11}, z_{21} \cdots z_{n1}\}), \min\{z_{12}, z_{22} \cdots z_{n2}\}), \cdots, \min\{z_{1m}, z_{2m} \cdots z_{nm}\}) \\ &= (z_1^-, z_2^- \cdots z_m^-) \end{aligned}$$

第二步，计算各指标与最优、最劣方案之间的距离D_i^+、D_i^-。W_{j}为第 j 个指标的权重，即重要程度。具体表达式如下：

$$D_i^+ = \sqrt{\sum_{j=1}^{m} W_j (z_j^+ - z_{ij})^2}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} W_j (z_j^- - z_{ij})^2}$$

第三步，计算各个指标与最佳方案的贴近程度M_i，M_i应该位于 0 和 1 之间，越接近 1，则表明该指标越优。具体表达式如下：

$$M_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

5.4.2 模型的求解

5.4.2.1 婴儿睡眠质量综合评判

运用 spearman 相关系数检验睡眠质量变量之间的相关性，得到整晚睡眠时间与睡醒次数之间的相关系数为-0.318，且 p=0.000<0.01，结果显著，说明二者之间存在强负相关；整晚睡眠质量与入睡方式之间的相关系数为 0.232，且 p=0.000<0.01，结果显著，说明二者之间存在强正相关；睡醒次数与入睡方式之间的相关系数为-0.255，且 p=0.000<0.01，结果显著，说明二者之间存在强负相关。

可以给出各睡眠质量变量之间的关系图：

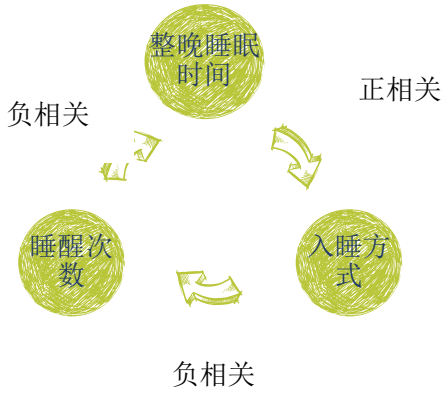


图 14 睡眠质量指标关系图

表 5.6 相关系数表

	整晚睡眠时间（时：分：秒）	睡醒次数	入睡方式
整晚睡眠时间（时：分：秒）	1(0.000***)	-0.318(0.000***)	0.232(0.000***)
睡醒次数	-0.318(0.000***)	1(0.000***)	-0.255(0.000***)
入睡方式	0.232(0.000***)	-0.255(0.000***)	1(0.000***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

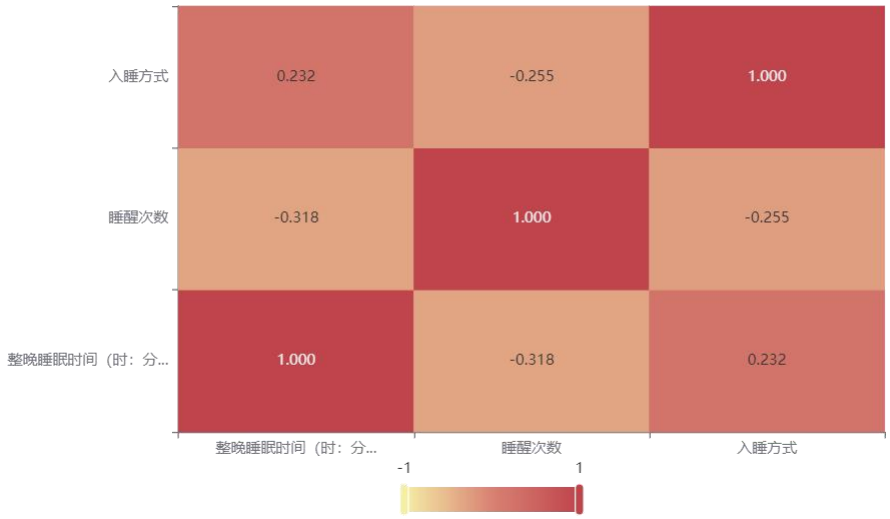


图 15 相关系数热力图

经过计算可以得到表中的权重结果，即整晚睡眠时间、睡醒次数、入睡方式分别占比为 0.297、0.419 和 0.284。结合之前的正负项变动，最终确定各婴儿睡眠质量 $Y_i(i=1,2,\cdots 399)$ 与各个睡眠质量指标 $Y_j(j=1,2,3)$ 间的评价函数关系式：

$$Z=W_1Y_1+W_2Y_2+W_3Y_3$$

即：

$$Z=0.297Y_1+0.419Y_2+0.284Y_3$$

表 5.7 权重表

项	指标变异性	指标冲突性	信息量	权重(%)
整晚睡眠时间（时：分：秒）	1.448	2.028	2.938	29.678
睡醒次数	1.622	2.56	4.152	41.949
入睡方式	1.403	2.001	2.808	28.372

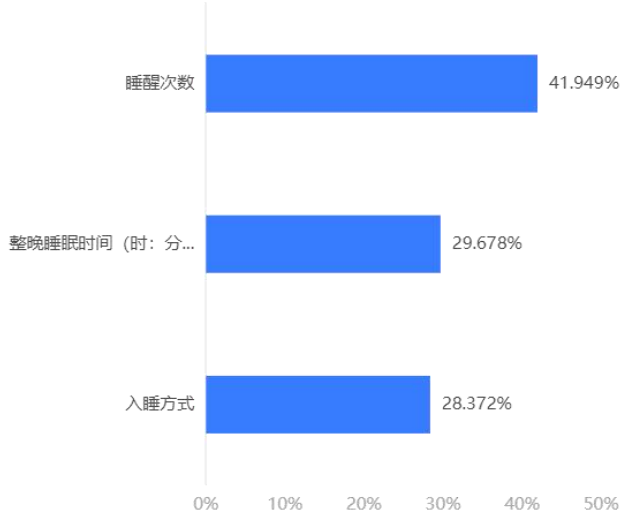


图 16 权重直方图

根据 5.4.1 的模型，通过 spss 可以得到 399 个婴儿样本睡眠质量得分并进行排序。然后运用 K-means 对综合评价分做聚类，分出优、良、中、差四类。结果如下所示：

表 5.8 评判等级结果

评判等级	区间下限
优	0.777871971989136
良	0.641337057538619
中	0.497799826945968
差	0

5.4.2.2 婴儿综合睡眠质量与母亲的身体、心理指标关联模型

（1）回归方程

对数据标准化处理后，用偏最小二乘法回归，得到回归方程如下所示：

$$整晚睡眠时间=0.761+0.001\times 母亲年龄+0.006\times CBTS-0.011\times EPDS+0.002\times HADS$$

$$睡醒次数=0.054+0.002\times 母亲年龄-0.005\times CBTS+0.006\times EPDS+0.001\times HADS$$

$$入睡方式=0.076-0.009\times 母亲年龄-0.008\times CBTS+0.01\times EPDS+0.01\times HADS$$

（2）最后 20 组婴儿综合睡眠质量预测

编号 391-410 婴儿的综合睡眠质量等级预测如表 5.9 所示：

表 5.9 评判等级结果

整晚睡眠时间	睡醒次数	入睡方式	等级
9.837	1.79	3.1	良

9.998	1.63	3.204	良
9.956	1.44	3.236	优
10.614	0.97	3.364	优
10.481	1.4	3.016	优
10.201	1.42	3	优
10.138	1.45	3.208	优
10.313	1.46	3.024	优
10.04	1.46	3.016	优
10.439	1.34	3.172	优
9.69	1.71	3.372	良
10.362	1.34	3.356	优
10.404	1.3	2.964	中
10.264	1.33	3.18	优
10.509	1.06	3.28	优
10.334	1.39	2.956	中
10.18	1.54	3.432	良
10.32	1.27	3.104	优
10.572	1.12	3.228	优
10.229	1.5	3.02	优

5.5 问题五的统计与求解

在第三问的基础上，引入新的约束条件，即使婴儿的睡眠质量为优，使用优化模型，构造三种心理指标治疗费用最低的目标函数，在第四问求解的婴儿综合睡眠质量 f 的评级基础上，列出新的约束条件：$f \geq 777871971989136$, 即婴儿睡眠质量为优。利用 Critic 模型确定婴儿整晚睡眠时间、睡醒次数和入睡方式的权重，分别约为 29%，28%和 41%，构建婴儿综合睡眠表达式：$f = 0.29x_{睡眠时间} + 0.28x_{入睡方式} + 0.41x_{睡醒次数}$

在此基础上，构建三种心理指标 CBTS、EPDS 和 HADS 的治疗费用w_3的目标函数：

$$\min w_3 = y_1 + y_2 + y_3$$

结合问题三和问题四列出以下约束条件：

$$\begin{cases} y_1 = 870.67x_1 + 200 \\ y_2 = 690x_2 + 500 \\ y_3 = 2440x_3 + 300 \\ y = 2.136 - 0.014x_{年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS} \\ x_{睡眠时间} = 10.528 - 0.043x_{CBTS} - 0.0789_{EPDS} + 0.014x_{HADS} \\ x_{入睡方式} = 2.898 + 0.027x_{CBTS} - 0.034x_{EPDS} + 0.083x_{HADS} \\ x_{睡醒次数} = 1.146 - 0.045x_{CBTS} + 0.052x_{EPDS} + 0.014x_{HADS} \\ f = 0.29x_{睡眠时间} + 0.28x_{入睡方式} + 0.41x_{睡醒次数} \\ f \geq 1.6987465 \\ y \leq 2 \\ x_{CBTS} = 15 - x_1 \\ x_{EPDS} = 22 - x_2 \\ x_{HADS} = 18 - x_3 \\ x_1, x_2, x_3 为整数 \end{cases}$$

利用 Lingo 进行最小二乘法求解，得到w=5830，即在第三问的基础上，使婴儿睡眠质量为优最少需要花费三种治疗费用 5830 元，需要进行七分的 EPDS 治疗，此时婴儿的行为特征为中等型。

在所求得的结果中，改变约束条件，令婴儿的行为特征转换为安静型，即$y \leq 1$，构建目标函数：

$$\min w_4 = y_1 + y_2 + y_3$$

列出约束条件：

$$\begin{cases} y_1 = 870.67x_1 + 200 \\ y_2 = 690x_2 + 500 \\ y_3 = 2440x_3 + 300 \\ y = 2.136 - 0.014x_{年龄} - 0.002x_{CBTS} + 0.008x_{EPDS} + 0.006x_{HADS} \\ x_{睡眠时间} = 10.528 - 0.043x_{CBTS} - 0.0789_{EPDS} + 0.014x_{HADS} \\ x_{入睡方式} = 2.898 + 0.027x_{CBTS} - 0.034x_{EPDS} + 0.083x_{HADS} \\ x_{睡醒次数} = 1.146 - 0.045x_{CBTS} + 0.052x_{EPDS} + 0.014x_{HADS} \\ f = 0.29x_{睡眠时间} + 0.28x_{入睡方式} + 0.41x_{睡醒次数} \\ f \geq 1.6987465 \\ y \leq 1 \\ x_{CBTS} = 15 - x_1 \\ x_{EPDS} = 22 - x_2 \\ x_{HADS} = 18 - x_3 \\ x_1, x_2, x_3 为整数 \end{cases}$$

利用 Lingo 进行最小二乘法求解，使婴儿的行为特征为安静型且综合睡眠质量为优的可行域不存在，即没有合适的解。

因此只有将婴儿行为特征控制在中等型，满足婴儿综合睡眠质量为优的三种心理指标 CBTS、EPDS 和 HADS 的最低治疗费用为 5830 元，需要进行七分的 EPDS 治疗。

六、模型检验

6.1 对问卷数据和相关性模型的检验

我们通过建立母亲心理指标和婴儿行为特征的柱状图，可视化了母亲心理指标和婴儿行为特征的变化趋势，初步验证了数据之间的相关性。从图 6-1 可以看出，母亲心理指标 CBTS（蓝色）、EPDS（绿色）、HADS（黑色）数值高的区域，婴儿行为特征（黄色）数值同样很高，并且 CBTS、EPDS、HADS 与婴儿行为特征呈现相同的变化趋势。以上分析验证了问卷数据的科学性和正确性，同时验证了模型的准确性。

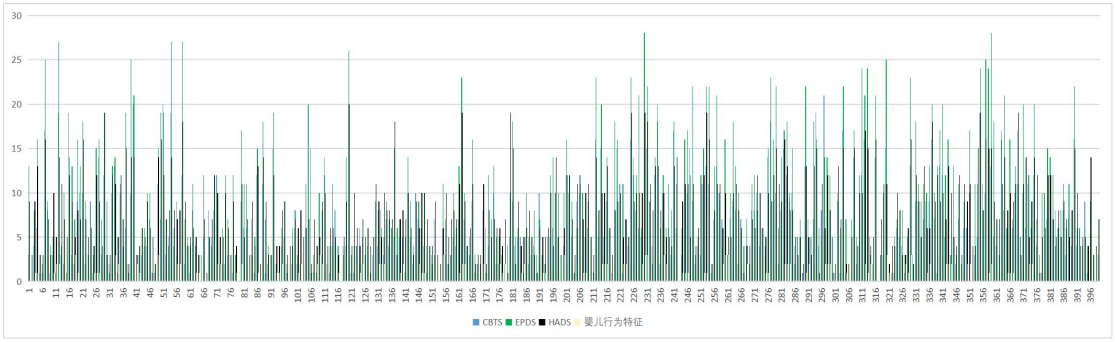


图 17 母亲心理指标和婴儿行为特征的柱状图

6.2 交叉验证：AdaBoost 分类模型

通过对 AdaBoost 进行五折交叉验证来检验模型准确率，如表 6.1 所示，交叉验证准确率为 0.53846154、0.51282051、0.56410256、0.57692308、0.58974359，平均准确率为 0.5564102564102564，证明所使用模型合理性和稳健性。

表 6.1 五折交叉验证检验结果

交叉验证准确率	平均准确率
0.53846154	0.5564102564102564
0.51282051	
0.56410256	
0.57692308	
0.58974359	

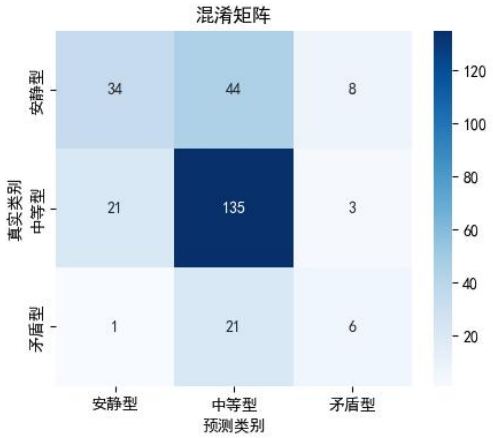


图 18 混淆矩阵检验图

七、模型评价

7.1 模型优点

- 数据分析处理：对原始数据进行了数据预处理工作，包括查找和剔除异常值，对未定类的数据进行编码，对心理指标的量表问卷进行信度和效度检验，即KMO 和巴特利特检验，提高了数据的准确性，便于后续相关模型的建立与问题求解。
- 皮尔逊相关性检验：快速准确地分析得到母亲身体指标和心理指标对于婴儿行为特征与睡眠质量是否存在相关关系和相关性，筛选出主要影响指标，便于量化分析各变量之间的关系。
- 多元线性回归模型：利用多元回归模型分析母亲身心健康指标与婴儿行为特征和睡眠质量之间的函数关系，分析各个变量对婴儿行为特征和睡眠质量的相关程度与拟合程度，简单便捷，便于预测未知项。
- 线性规划模型：根据题目已知条件和求解出的母亲身心健康指标与婴儿行为特征、睡眠质量的函数关系式构建约束条件，可以求解出最优的治疗方案实现婴儿行为特征的转变和睡眠质量的提高，具有很强的实用价值。
- 多分类模型：Adaboost 是一种有很高精度的分类器，可以使用各种方法构建子分类，构建弱分类器操作简单，减少特征筛选步骤，较为简单便捷地实现婴儿行为特征分类。
- critic 权重法模型：属于客观赋权法，根据各项指标值的变异程度确定权数，一定程度上避免了主观性带来的偏差，更好地确定婴儿整晚睡眠时间、入睡方式和睡醒次数的权重。
- K-means 算法：该算法具有优化迭代功能，能根据较少的已知聚类样本的类别对树进行剪枝确定部分样本的分类，针对本题样本数量较小和已知分类数量的情况下可以降低总分类时间的复杂度。

7.2 模型缺点

- 数据质量：由于数据是由问卷等调查统计而来，数据的准确性和可靠性依赖于问卷填写的质量，包括填写数据时的有效性和真实性，填写答卷时调查对象的主观性也会对数据的准确性造成一定影响，进而影响建模和模型求解的结果。
- 主观性：在模型选择方面由于研究者主观判断和个人爱好偏向，针对同一问题可能会选择不同模型，计算结果也会有所差异。在利用 K-means 算法时，需根据初始聚类中心来确定一个初始划分，进而对初始划分进行优化。对初始值的选择的具有一定主观性，对聚类结果造成一定程度的影响。
- 依赖假设：模型的可靠性一定程度以来假设的合理性，比如问卷数据在剔除异常值和进行数据预处理之后是准确的，以及婴儿行为特仅受母亲身体指标和心理指标的影响，婴儿的睡眠质量仅受母亲的身心健康指标、整晚睡眠时间、入睡方式和睡醒次数影响。

7.3 模型推广与改进

- 对于数据类型可从问卷调查数据扩展至观测数据、实验数据等其他来源和类型的数据，实验数据通过反复多次的实验得到，科学性强，相较问卷数据可在一

定程度上改善问卷数据存在填写不准确、不规范和主观性强的问题，增加数据样本，提升模型预测的结果的准确性。同时，在问卷调查中，应增加问卷调查对象，扩展样本容量，避免数据太少影响研结果准确性的问题。

- 对于使用范围的推广，论文中使用的各类模型不仅可以用于母亲身心健康对婴儿成长影响这一社会、心理学领域，还适用于经济、体育、医学等不同领域，选取合适的模型解决更多类型的问题，发挥模型的理论和实际意义。
- 对于综合评价模型的推广，可以在问卷中增加更广泛的多类型问题，引入多种不同类型的可能的影响因素，对综合模型建立时所考虑的评价指标更全面、多样，有助于提高评价的准确性。此外，在推广模型时要进行模型的适用性检验。

八、参考文献

[1]刘卓娅,郭玉琴,宋娟娟等.婴幼儿入睡方式及其对睡眠质量的影响[J].中国当代儿科杂志,2022,24(03):297-302.

[2]王念蓉,叶亚.婴儿夜晚睡眠模式发展的前瞻性研究[J].中国当代儿科杂志,2016,18(04):350-354.

[3]梁北辰,戴景民.偏最小二乘法在系统故障诊断中的应用[J].哈尔滨工业大学学报,2020,52(03):156-164.

[4]DIAKOULAKI D, MAVROTAS G, PAPAYANNAKIS L. Determining objective weights in multiple criteria problems: The CRITIC method [J]. Computer Ops Res, 1995, 22: 763-770.

[5]张羽頔,祁月,马雪梅等.母亲产后抑郁症状与 1.5～2 月龄婴儿发育的关联分析[J].中国妇幼健康研究,2020,31(07):889-894.

[6]石影,张永花,张婧洁等.母亲产后抑郁对婴儿体格和神经心理发育的影响[J].甘肃医药,2022,41(12):1067-1069.

[7]张小甜,张悦,邓梁琼等.母亲抑郁情绪对 3 月龄婴儿发育影响的关联性分析[J].中国儿童保健杂志,2022,30(09):947-951.

九、附录

第 2 问：

基于 AdaBoost 集成学习算法的婴儿行为特征分类模型

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import confusion_matrix
```

```
df = pd.read_excel('c.xlsx')
df['behavior'].values
X = df.loc[:,['m_age','marriage','edu','pregnancy_time','deliver','CBTS','HADS','EPDS']]
y = df['behavior']

# 拆分训练集和测试集，最优训练/测试 0.7  最佳随机种子_1 为 2225
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.7, random_state=2225)
# 创建 AdaBoost 分类器,基分类器为决策树，最优基分类器个数 4
base_classifier = DecisionTreeClassifier(max_depth=4)
#最佳迭代次数 67  最佳学习率 1.0  最佳随机种子_2 为 42
adaboost_clf = AdaBoostClassifier(base_classifier, n_estimators=67, learning_rate=1.0, random_state=42)

adaboost_clf.fit(X_train, y_train)
y_pred = adaboost_clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("准确率：", accuracy,y_pred)

kf = KFold(n_splits=5, shuffle=True, random_state=42)
cv_scores = cross_val_score(adaboost_clf, X, y, cv=kf)
print("交叉验证准确率：", cv_scores)
print("平均准确率：", cv_scores.mean())

cm = confusion_matrix(y_test, y_pred)
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
class_labels = ['安静型', '中等型', '矛盾型']
plt.figure(figsize=(5, 4))
sns.heatmap(cm,annot=True,fmt="d",cmap="Blues",xticklabels=class_labels,yticklabels=class_labels)
plt.xlabel("预测类别")
plt.ylabel("真实类别")
plt.title("混淆矩阵")
plt.show

df = pd.read_excel('q2_4.xlsx')
XX = df.loc[:,['m_age','marriage','edu','pregnancy_time','deliver','CBTS','HADS','EPDS']]
yy = adaboost_clf.predict(XX)
print(yy)
```

第 3 问：

多元线性回归模型

model:

min=y1+y2+y3;

```
y1=870.67*x1+200;
y2=690*x2+500;
y3=2440*x3+300;
    y=2.136-0.014*24-0.002*(15-x1)+0.008*(22-x2)+0.006*(18-x3);
    f=0.29*X1+0.41*X2+0.28*X3;
 f>=0.777871971989136;
   Xa=10.528+0.043*(15-x1)-0.079*(22-x2)+0.014*(18-x3);
Xb=1.146-0.045*(15-x1)+0.052*(22-x2)+0.014*(18-x3);
Xc=2.898+0.027*(15-x1)-0.034*(22-x2)+0.038*(18-x3);

y<=2;
x1<=15;
x2<=22;
x3<=18;
@gin(x1);@gin(x2);@gin(x3);
end
```

Global optimal solution found.

Objective value:	5830.000
Objective bound:	5830.000
Infeasibilities:	0.000000
Extended solver steps:	0
Total solver iterations:	2
Elapsed runtime seconds:	0.95

Model Class:	MILP

Total variables:	7
Nonlinear variables:	0
Integer variables:	3

Total constraints:	6
Nonlinear constraints:	0

Total nonzeros:	14
Nonlinear nonzeros:	0

Variable	Value	Reduced Cost
Y1	200.0000	0.000000
Y2	5330.000	0.000000
Y3	300.0000	0.000000

X1	0.000000	870.6700
X2	7.000000	690.0000
X3	0.000000	2440.000
Y	1.998000	0.000000

Row	Slack or Surplus	Dual Price
1	5830.000	-1.000000
2	0.000000	-1.000000
3	0.000000	-1.000000
4	0.000000	-1.000000
5	0.000000	0.000000
6	0.2000000E-02	0.000000

Global optimal solution found.

Objective value:	92080.00
Objective bound:	92080.00
Infeasibilities:	0.000000
Extended solver steps:	0
Total solver iterations:	2
Elapsed runtime seconds:	0.09
Model Class:	MILP
Total variables:	7
Nonlinear variables:	0
Integer variables:	3
Total constraints:	6
Nonlinear constraints:	0
Total nonzeros:	14
Nonlinear nonzeros:	0

Variable	Value	Reduced Cost
Y1	200.0000	0.000000
Y2	91580.00	0.000000
Y3	300.0000	0.000000
X1	0.000000	870.6700
X2	132.0000	690.0000
X3	0.000000	2440.000
Y	0.9980000	0.000000

Row	Slack or Surplus	Dual Price
1	92080.00	-1.000000
2	0.000000	-1.000000
3	0.000000	-1.000000
4	0.000000	-1.000000
5	0.000000	0.000000
6	0.2000000E-02	0.000000

第 4 问：

Spearman 相关性分析

```
import pandas
from spsspro.algorithm import descriptive_analysis
data = df.loc[:,[‘sleep_time’,’wake_times’,’sleep_way’]]
result = descriptive_analysis.correlation_analysis(data)
print(result)
```

TOPSIS 综合评价

```
import numpy
import pandas
from spsspro.algorithm import quantify_analysis

data = df.loc[:,[‘sleep_time’]]
forward = df.loc[:,[’wake_times’]]
reverse = df.loc[:,[’sleep_way’]]

weight = pandas.Series([0.497, 0.42, 0.283], name="D")
print(quantify_analysis.topsis_analysis(data, forward, reverse, index, weight))
```

K 聚类分析代码

```
import numpy
import pandas
from spsspro.algorithm import statistical_model_analysis
data = df.loc[:,[‘score’]]
result = statistical_model_analysis.cluster_analysis(data, cluster_num=3)
print(result)
```

利用偏最小二乘法回归方程解睡眠指标

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score
```

```
from sklearn.model_selection import cross_val_score, KFold

df = pd.read_excel('c_4.xlsx')

# df['behavior'].values

X = df.loc[:,['m_age','CBTS','HADS','EPDS']]
# X_4 = df.loc[:,['m_age','CBTS','HADS','EPDS']]
y_1 = df['sleep_time']
y_2 = df['wake_times']
y_3 = df['sleep_way']

df_q4 = pd.read_excel('q2_4.xlsx')

df_age = df_q4.loc[:,['m_age']]
df_CBTS = df_q4.loc[:,['CBTS']]
df_EPDS = df_q4.loc[:,['EPDS']]
df_HADS = df_q4.loc[:,['HADS']]

#整晚睡眠时间预测
def calculate_time(age, CBTS, EPDS, HADS):
    t = 0.761 + 0.001 * age + 0.006 * CBTS - 0.011 * EPDS + 0.002 * HADS
    return t
age = df_age.values.ravel()
CBTS = df_CBTS.values.ravel()
EPDS = df_EPDS.values.ravel()
HADS = df_HADS.values.ravel()
results_1 = []
for i in range(len(age)):
    t = calculate_time(age[i], CBTS[i], EPDS[i], HADS[i])
    results_1.append(t)
for i, result_t in enumerate(results_1):
    print(f"t{ i+391 } =", result_t)

#睡醒次数预测
def calculate_num(age, CBTS, EPDS, HADS):
    n = 0.054 + 0.002 * age - 0.005 * CBTS + 0.006 * EPDS + 0.001 * HADS
    return n
age = df_age.values.ravel()
CBTS = df_CBTS.values.ravel()
EPDS = df_EPDS.values.ravel()
HADS = df_HADS.values.ravel()
results_2 = []
for i in range(len(age)):
```

```
    num = calculate_num(age[i], CBTS[i], EPDS[i], HADS[i])
    results_2.append(num)
for i, result_num in enumerate(results_2):
    print(f"t{ i+391 } =", result_num)

#入睡方式预测
def calculate_way(age, CBTS, EPDS, HADS):
    w = 0.76 - 0.009 * age + 0.008 * CBTS - 0.01 * EPDS + 0.01 * HADS
    return w
age = df_age.values.ravel()
CBTS = df_CBTS.values.ravel()
EPDS = df_EPDS.values.ravel()
HADS = df_HADS.values.ravel()
results_3 = []
for i in range(len(age)):
    way = calculate_way(age[i], CBTS[i], EPDS[i], HADS[i])
    results_3.append(way)

for i, result_way in enumerate(results_3):
    print(f"t{ i+391 } =", result_way)

  list = np.array([[]])
def sleep_func(t,n,w):
    time = t * (12 - 5) + 5
    number = n * (10 - 0) + 0
    way = w * (5 - 1) + 1
    return time,number,way

print('391-410 号样本整晚睡眠时间：')
for n in range(0,20):
    time_s = results_1[n]
    num_s = results_2[n]
    way_s = results_3[n]
    time_o,num_s,way_s = sleep_func(time_s,num_s,way_s)
    print(time_o)
print('391-410 号样本睡醒次数：')
for n in range(0,20):
    time_s = results_1[n]
    num_s = results_2[n]
    way_s = results_3[n]
    time_o,num_s,way_s = sleep_func(time_s,num_s,way_s)
    print(num_s)
print('391-410 号样本入睡方式：')
for n in range(0,20):
```

```
    time_s = results_1[n]
    num_s = results_2[n]
    way_s = results_3[n]
    time_o,num_s,way_s = sleep_func(time_s,num_s,way_s)
    print(way_s)
```

#基于 KNN 的睡眠质量分类模型

```
X_sleep = df.loc[:,['sleep_time','wake_times','sleep_way']]
y_sleep = df['quality']

list = [0.01]
max_acc = list[0]



# 将数据集划分为训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X_sleep, y_sleep, test_size=0.8, random_state=1804)

# 创建 K 近邻分类模型，指定 K 值为 3
knn = KNeighborsClassifier(n_neighbors=6)

# 训练模型
knn.fit(X_train, y_train)

# 使用训练好的模型进行预测
y_pred = knn.predict(X_test)

# 计算模型准确度
accuracy = accuracy_score(y_test, y_pred)
print("模型准确度：", accuracy)

XX_sleep = df_q4.loc[:,['sleep_time','wake_times','sleep_way']]
yy_pred = knn.predict(XX_sleep)
for n in range(0,20):
    print(yy_pred[n])
```

第 5 问

婴儿行为特征为中等型时

model:

```
    min=y1+y2+y3;

    y1=870.67*x1+200;
    y2=690*x2+500;
    y3=2440*x3+300;
```

```
        y=2.136-0.014*24-0.002*(15-x1)+0.008*(22-x2)+0.006*(18-x3);
        f=0.29*X1+0.41*X2+0.28*X3;
      f>=0.777871971989136;
       Xa=10.528+0.043*(15-x1)-0.079*(22-x2)+0.014*(18-x3);
    Xb=1.146-0.045*(15-x1)+0.052*(22-x2)+0.014*(18-x3);
    Xc=2.898+0.027*(15-x1)-0.034*(22-x2)+0.038*(18-x3);

    y<=2;
    x1<=15;
    x2<=22;
    x3<=18;
    @gin(x1);@gin(x2);@gin(x3);
end
```

婴儿行为特征为安静型时

```
model:
    min=y1+y2+y3;

    y1=870.67*x1+200;
    y2=690*x2+500;
    y3=2440*x3+300;
        y=2.136-0.014*24-0.002*(15-x1)+0.008*(22-x2)+0.006*(18-x3);
        f=0.29*X1+0.41*X2+0.28*X3;
      f>=0.777871971989136;
       Xa=10.528+0.043*(15-x1)-0.079*(22-x2)+0.014*(18-x3);
    Xb=1.146-0.045*(15-x1)+0.052*(22-x2)+0.014*(18-x3);
    Xc=2.898+0.027*(15-x1)-0.034*(22-x2)+0.038*(18-x3);

    y<=1;
    x1<=15;
    x2<=22;
    x3<=18;
    @gin(x1);@gin(x2);@gin(x3);
end
```