所属类别	2023 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科组		CM2304403

基于逻辑回归模型探究母亲身心健康对婴儿成长影响

摘　要

母亲扮演婴儿生命中最重要的角色之一，身心健康不良的母亲可能会对婴儿的认知、性格、行为习惯等方面产生负面影响。本文从母亲的身心健康和婴儿的行为习惯出发，基于逻辑回归模型探究母亲身心健康对婴儿成长的影响。

针对问题一，本文对母亲的 CBTS、EPDS、HADS 与年龄这四种连续型变量进行**正态分布检验**。检验后，按照变量的数值表现分别使用卡方检验、单因素方差分析和斯皮尔曼相关系数分析，再对母亲与婴儿各指标之间的变量进行**相关性-显著性检验**。数据体现出，母亲的身体指标对婴儿的行为特征无显著影响，对睡眠质量有轻微影响。母亲的心理指标对婴儿的行为特征和睡眠质量有较大影响，其中对睡眠质量的影响较为显著。

针对问题二，婴儿行为特征是离散型变量，符合**随机森林算法**的特征，故对前 250 个样本数据进行随机森林测试组训练，再对剩余样本进行检测，最终得到准确率为 52%的随机森林模型。考虑到样本不均对随机森林模型准确性的影响，本文对婴儿的行为特征进行**欠采样**和**重采样**，利用 Excel 的随机数生成函数重新刷新数据次序以排除干扰，最终获得总体准确率达到 54%的随机森林模型。为得到更为准确的预测模型，基于问题一，本文对母亲身心指标的多个变量主成分分析进行降维，随后进行**多元逻辑回归**，得到准确率为 70.3%的多元 Logistic 回归模型进行再预测，最终获得二十个婴儿的预测数据。结果可见下文表六。

针对问题三，基于问题二所得的多元逻辑回归模型，以治疗病症最大化、费用最小为目标，使用**穷举法**列举三项指标降低的可能性，用随机森林模型判断当时婴儿的行为特征。最终在 6992 种方案中，本文计算出将 238 号婴儿行为特征由矛盾型转为中等型的最低治疗费用为 3085 元。将行为特征变为安静型的最优方案需将 CBTS 降低 1 分，HADS 降低 2 分，治疗费用为 6750.67 元。

针对问题四，首先对指标进行类型方向统一和 **Min-Max 标准化**处理，所有数值收敛至 0-1 区间。数据预处理完成后，建立**秩和比综合评价体系**，并通过由 probit 值计算得到的回归方程对婴儿的综合睡眠质量进行分类。最后根据已获得的睡眠质量数据和对应母亲的身体和心理指标，将数据再次代入到**多元逻辑回归**模型中进行训练和测试，分析得到最后 20 婴儿的综合睡眠质量，结果可见下文表八。

针对问题五，本文在问题三建立的穷举数据表上筛选出中等型和安静型的治疗方案，在问题四建立的多元逻辑回归模型中筛选出睡眠综合质量为优的，对其心理指标治疗结果数据反向标准化和正向化，使其还原为原始得分。得分取交集后带入问题三中得出的**线性方程**计算出最终费用并排序筛选，获得在中等型基础上最少费用为 6915 元，在安静型基础上最少费用为 10046.34 元。

关键词：相关性-显著性检验　随机森林　多元逻辑回归　穷举法　秩和比综合评价

一、 问题重述

母亲扮演婴儿生命中最重要的角色之一，她既给婴儿提供营养与保护，又给予婴儿极大的情感支持和安全感。母亲的身心状况可能会影响婴儿的生理和心理发展，身心健康不良的母亲可能会对婴儿的认知、性格、行为习惯等方面产生负面影响。附件展示了 390 名婴儿及其母亲的身体指标等相关数据，表 1 展示了患病得分与治疗费用的关系，现要求通过建立数学模型完成下列问题：

问题一：根据附件数据探究母亲身体和心理相关指标对婴儿的行为特征和睡眠质量的影响，研究其规律。

问题二：建立婴儿的行为特征与母亲的身体指标和心理指标的关系模型，并完成附件表最后 20 组婴儿行为特征信息空缺的填写。

问题三：已知患病程度的变化率与治疗费用皆成正比，通过表 1 患病得分与治疗费用的关系建立模型，分析将编号为 238 的婴儿行为特征由矛盾型转为中等型的最低治疗费用。若将其行为特征转为安静型，治疗方案需如何调整。

问题四：将婴儿的睡眠质量分成优、良、中、差四个评级，并建立婴儿综合睡眠质量与母亲身体心理指标的关联模型，完成附件表最后 20 组婴儿综合睡眠质量信息空缺的填写。

问题五：在问题三的基础上，调整治疗策略使 238 号婴儿的睡眠质量评级为优。

二、 问题分析

2.1 问题一的分析

问题一要求通过附件中的数据分析母亲的身体指标和心理指标对婴儿行为特征和睡眠特征的影响。首先对数据进行预处理，通过单因素方差分析探求睡眠时间、婚姻状况与各项指标的相关性，用多元线性回归算法将睡眠时间异常值转化，用二元逻辑回归算法将婚姻状况异常值转化，并依次进行了数据的正向化与标准化处理。母亲的身体指标可分为母亲年龄、婚姻状况、妊娠时间、分娩方式、教育程度五类，心理指标可由爱丁堡产后抑郁量表（EPDS）、医院焦虑抑郁量表（HADS）、分娩相关创伤后应激障碍（CBTS）得分正向化、归一化处理后进行分析，最后通过判断相关变量的连续与离散类型，用斯皮尔曼相关系数分析计算连续变量之间的相关系数，用卡方检验计算类别变量与类别变量之间的显著性，用单因素方差分析计算多分类变量和连续变量之间的显著性，从而得出母亲身体指标、心理指标的各小类别对婴儿行为特征和睡眠质量的影响，得出结论。

2.2 问题二的分析

问题二要求建立婴儿的行为特征与母亲的身体、心理指标的关系模型。考虑到问题二是婴儿行为特征的分类任务，其中母亲的身体指标中许多非线性特征如婚姻状况、教育程度、分娩方式有利于决策树模型的创建，故利用基于决策树的随机森林模型建立数据的训练集和测试集，得到随机森林的预测模型，对最后 20 组(编号 391-410 号)婴儿的行为特征进行预测。同时为了进一步提高结果的准确性，在问题一斯皮尔曼系数表反映婴儿的行为特征与母亲的心理指标相关性较显著而与母亲的身体指标相关

性较不显著的基础上，我们需要通过对母亲的身体指标、心理指标进行主成分分析，计算其贡献率对多种变量进行降维，以确定行为特征与更高贡献率的指标构建回归模型。完成降维后建立有监督模型对母亲身体、心理指标和婴儿行为特征进行多元逻辑回归，得到以母亲的身体、心理指标为 X，以婴儿的行为特征为 Y 的回归预测模型，基于上述随机森林预测模型和多分类逻辑回归模型，得到最后 20 组(编号 391-410 号)婴儿行为特征的预测结果并进行对比分析。

2.3 问题三的分析

问题三要求我们建立一个治疗费用与 CBTS、EPDS、HADS 三项指标分数治疗的模型，并求出编号为 238 的婴儿行为特征由矛盾型转为中等型所需的最低治疗费用以及将其行为特征转化为安静型的治疗方案。由题目提供信息可得，CBTS、EPDS、HADS 的治疗费用相对于患病程度呈现一个正向线性关系，故治疗费用与三者呈现正相关。由此可知我们可以建立一个由 CBTS、EPDS、HADS 为自变量，治疗费用为因变量的线性规划方程组，在问题二所得的婴儿行为特征所对应的三项指标的分数区间基础上，我们得知当婴儿的行为特征从矛盾型降低至中等型，各项指标的分数都会降低。通过穷举法列举各项心理问卷指标降低分数产生得到可能性，及可由线性规划方程组算出最低治疗费用。同理可得到婴儿的行为特征降低到安静型所需要的分数。

2.4 问题四的分析

问题四要求依据整晚睡眠时间、睡醒次数、入睡方式这三个指标将婴儿的睡眠质量进行优、良、中、差四个等级的分类，并因此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。我们可以首先建立一个分类模型，通过秩和比综合评价法计算 RSR（秩和比值），将儿童的综合睡眠质量分为优、良、中、差四类。建立综合睡眠质量与母亲身体指标、心理指标的多元逻辑回归模型，将这 20 组婴儿母亲的身体指标和心理指标代入到上述模型中，及可得到对应的预测值。

2.5 问题五的分析

问题五要求在问题三的基础上调整治疗策略使 238 号婴儿的睡眠质量评级为优。根据问题四逻辑回归可得以母亲的身体、心理指标与婴儿的睡眠质量的逻辑回归预测模型，求出睡眠质量为优的母亲心理指标，对得到的心理指标治疗结果数据进行反向的标准化和正向化，使其还原为原始数据得分，最后用问题三中所得的心理指标与治疗费用线性规划方程组，得到调整后的治疗策略。

三、 模型假设

1. 母亲心理指标中的三个数值相互独立，互不干扰，在心理治疗时可以组合治疗而不会互相影响分数。

2. 每一次治疗的起步价为 1000 元，不考虑实际情况中拒付少付的情况。

3. 婴儿的行为特征能够通过母亲后天的身体和心理状况改善而发生改变。

4. 治疗方案以降低至婴儿行为特征转变即可，而不用将分数降为 0。

5. 仅考虑母亲的身体和心理特征对婴儿的行为特征影响，而需要排除环境因素或者父亲对孩子行为特征的影响。

四、 符号说明

符号	说明
Z	对数据进行标准化处理后的结果
P	正态分布检验中的的显著性水平用于反映正态分布情况
β	置信水平，描述在该置信水平下估计结果是否可靠，本文以 95%和 90%为准
R	协方差矩阵，用于衡量变量之间的相关性，本文中用其判断八项指标之间的相关性
$\hat{y}_i$	多分类多元逻辑回归预测模型输出值，各个类别值之间取最高的概率输出对应逻辑值
$\hat{\beta}_i$	逻辑回归模型中的回归系数或参数的估计值，衡量了每一个自变量对因变量的影响程度
x_i	表示回归模型中所有自变量的取值，可以为连续性变量也可以为离散型变量
RSR	秩和比值，本文用其判断睡眠数据的小项对于整体睡眠质量的影响程度

五、 模型的建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

首先筛选出数据异常值。针对整晚睡眠时间为“99:99”的异常值，用多元线性回归法探究整晚睡眠时间与睡醒次数的关系，并用多元逻辑回归模型探究整晚睡眠时间与入睡方式、婴儿行为特征的关系。由最终通过该异常值的睡醒次数、入睡方式和婴儿行为特征，将其转化为“12：00：00”；针对婚姻状况中数值为“3”“6”的异常值，通过与母亲年龄进行多元线性回归分析，将其皆转化为“2”，即“已婚”。

完成异常值处理后，用睡醒次数、CBTS、EPDS、HADS 分数各项极小型指标变量变量的实际最大值减去实际值进行正向化处理获得新的数据，完成对指标方向统一。同时我们将入睡方式从类别变量转换为数值变量且评判为极大型指标，将字符串变量“婴儿的行为特征”改为数字（矛盾型为 1，中等型为 2，安静型为 3）。

完成上述步骤后，将每一列的母亲身体和心理数据导入 MATLAB 构建矩阵 X，进行数据标准化处理获得矩阵 Z。标准化处理公式如下：

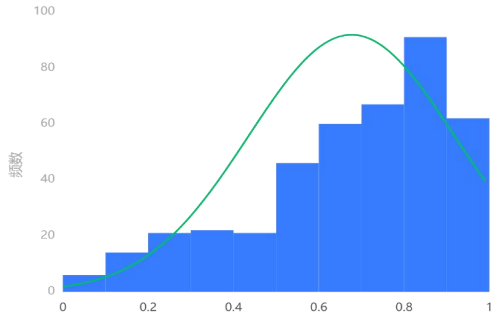
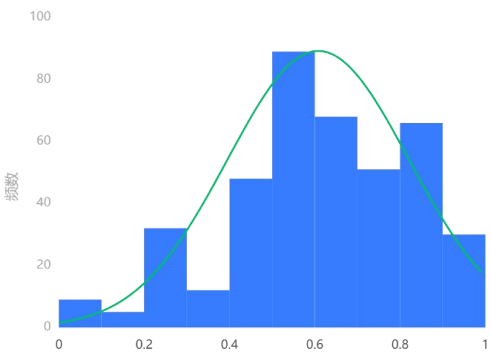
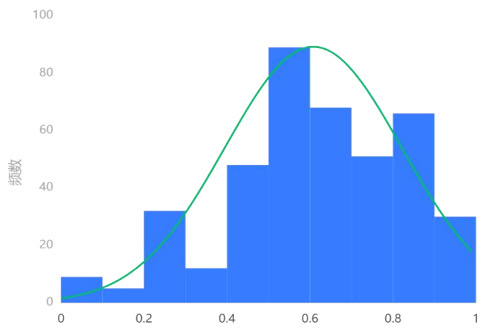
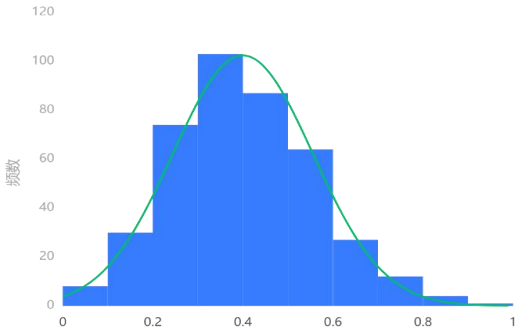
$$Z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \quad (1)$$

其中Z_{ij}表示将每一列中每一行X_{ij}的数据除以该列所有数据的平方和$\sqrt{\sum_{i=1}^{n} x_{ij}^2}$

5.1.2 正态分布检验

对于附件二中的重点数据，需要对连续性变量来验证其是否符合正态分布。我们对连续性数据使用 Jarque-Bera 检验其显著性 P，与 0.05 进行比较，若 P>0.05，则符合正态分布，若 P<0.05，则不符合正态分布。结合曲线直方图、Q-Q 图可以进一步直观反映是否符合正态分布。通过对 CBTS、EPDS、HADS 与母亲年龄样本进行分析，得到如下图所示的正态性检验直方图。其中 CBTS、EPDS 虽然没有呈现出绝对正态分布，但是基本呈现出钟形，基本也可以理解为符合正态分布。

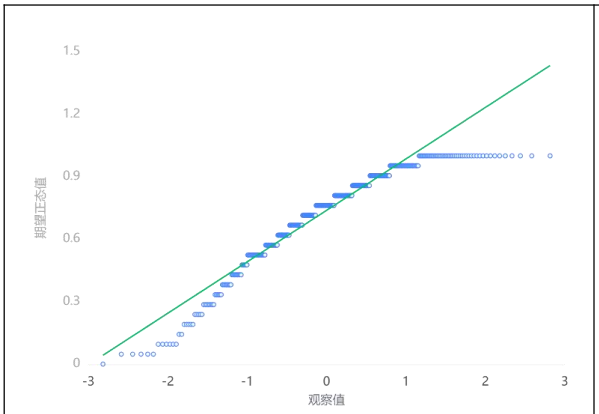
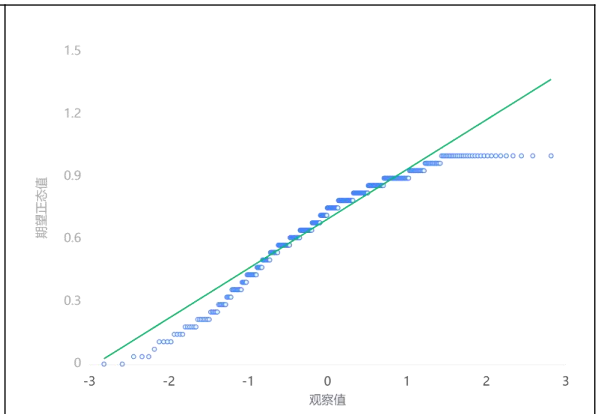
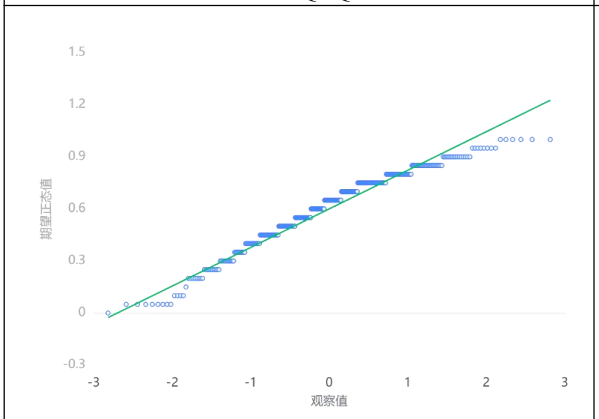
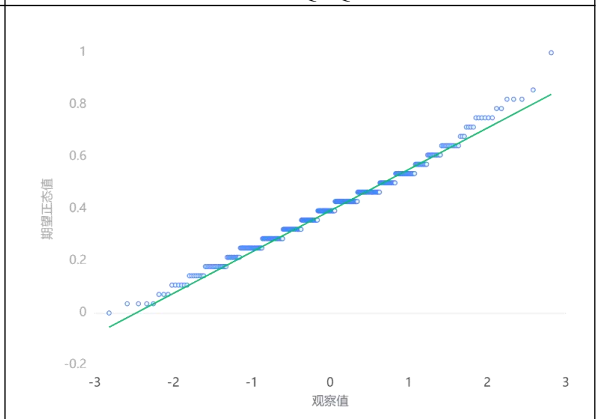
图一：正态分布检验图

<table>
<tr><td>  </td><td>  </td></tr>
<tr><td>CBTS 正态分布检验图</td><td>EPDS 正态分布检验图</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>HADS 正态分布检验图</td><td>母亲年龄正态分布检验图</td></tr>
</table>

5.1.3 Q-Q 图检验

以下 Q-Q 图以图形的方式将观测值与预测值（正态分布下的预测值），以实际值为横坐标，预测值为纵坐标绘制，散点若与直线拟合度越高则越服从正态分布。从以下 Q-Q 图可知四项数据大致通过正态分布检验。

图二：Q-Q 图检验

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>CBTS Q-Q 检验图</td><td>EPDS Q-Q 检验图</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>HADS Q-Q 检验图</td><td>母亲年龄 Q-Q 检验图</td></tr>
</table>

5.1.4 相关性—显著性检验

建立婴儿的行为特征和睡眠质量模型，首先需要对各个指标进行属性的分类。本题中，婚姻状况、教育程度、分娩方式、婴儿行为特征、婴儿性别、入睡方式为离散型指标，母亲年龄、妊娠时间、CBTS、EPDS、HADS、婴儿年龄、整晚睡眠时间、睡醒次数为连续型指标。将连续-连续指标进行斯皮尔曼相关系数分析，将连续-离散指标进行单因素方差分析，将离散-离散指标进行卡方检验分析，通过 SPSS 运算，获得例如下表的多个三种类型的表格：

斯皮尔曼检验示例表：

表一：斯皮尔曼相关性分析检验

		CBTS	EPDS	HADS	睡醒次数	整晚睡眠时间（时：分：秒）
CBTS	相关系数	1	.780**	.704**	0.08	.117*
	Sig.（双尾）	.	0	0	0.116	0.018
	N	410	410	410	390	410
EPDS	相关系数	.780**	1	.782**	.117*	.147**
	Sig.（双尾）	0	.	0	0.02	0.003
	N	410	410	410	390	410
HADS	相关系数	.704**	.782**	1	0.069	.104*
	Sig.（双尾）	0	0	.	0.171	0.036
	N	410	410	410	390	410
睡醒次数	相关系数	0.08	.117*	0.069	1	.312**
	Sig.（双尾）	0.116	0.02	0.171	.	0
	N	390	390	390	390	390
整晚睡眠时间（时：分：秒）	相关系数	.117*	.147**	.104*	.312**	1
	Sig.（双尾）	0.018	0.003	0.036	0	.
	N	410	410	410	390	410

** 在 0.01 级别（双尾），相关性显著。

* 在 0.05 级别（双尾），相关性显著。

单因素 ANOVA 检验示例表：

表二：入睡方式和 CBTS、EPDS、HADS 的单因素 ANOVA 检验

		平方和	自由度	均方	F	显著性
CBTS	组间	0.608	4	0.152	2.712	0.03
	组内	21.577	385	0.056		
	总计	22.185	389			
EPDS	组间	0.429	4	0.107	1.817	0.125
	组内	22.745	385	0.059		
	总计	23.174	389			
HADS	组间	0.253	4	0.063	1.376	0.241
	组内	17.7	385	0.046		
	总计	17.953	389			

卡方检验示例表：

表三：分娩方式和入睡方式卡方检验

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	2.094a	4	0.718
似然比	3.396	4	0.494
线性关联	1.531	1	0.216
有效个案数	390		

a 5 个单元格（50.0%）的期望计数小于 5。最小期望计数为 .26。

将表格所需信息进行组合整理，获得如下表：

表四：母亲身体指标、心理指标与婴儿行为方式、睡眠质量关系表

	婚姻状况	分娩方式	教育程度	母亲年龄	妊娠时间（周数）	CBTS	EPDS	HADS
婴儿行为特征	0.829	0.65	0.545	0.07	0.787	0.931	0.635	0.734
入睡方式	0.071	0.16	0.025	-0.081(0.110)	0.053(0.300)	-0.044(0.389)	-0.004(0.932)	-0.057(0.260)
睡醒次数	0.439	0.78	0.108	-0.038(0.450)	-0.087(0.088*)	0.08(0.116)	0.117(0.020**)	0.069(0.171)
整晚睡眠时间（时：分：秒）	0.666	0.045	0.653	0.013(0.805)	0.078(0.123)	0.132(0.009***)	0.17(0.001***)	0.122(0.016**)

备注：	
	由卡方检验算出的渐进显著性
	由单因素方差分析算出的显著性
	斯皮尔曼相关系数

对斯皮尔曼相关系数进行分析，$|\rho|$越接近 1，两个变量之间的相关性越强。由此可得，就整晚睡眠时间而言，CBTS、EPDS、HADS 对其的影响较为显著，即母亲的心理指标与整晚睡眠时间关系较显著，其中 EPDS 与其的相关系数绝对值最大，为 -0.166，关系最为显著；母亲的身体指标，如母亲年龄、妊娠时间则对其没有显著影响。就睡眠次数而言，EPDS 与其的相关系数绝对值最大，为 0.117，对其的影响较为显著，其余指标对其无显著影响。

对由卡方检验算出的渐进显著性和单因素方差分析算出的显著性进行分析，若其显著性 $p<0.05$，则可判断为二者之间具有显著差异，即二者相关性较显著，且 p 越小，二者相关性越强。由此可得，母亲的婚姻状况、分娩方式、教育方式与婴儿的入睡方式、婴儿行为特征关系、整晚睡眠时间、睡醒次数皆较为不显著；CBTS 对入睡方式有较显著影响；EPDS 对婴儿行为特征有较显著影响。将置信水平β调至 90%，可得教育程度与婴儿行为特征有轻微相似性。

综上，母亲的身体指标对婴儿的行为特征无显著影响，对睡眠质量有轻微影响。母亲的心理指标对婴儿的行为特征和睡眠质量均有影响，其中对睡眠质量的影响较为显著。

5.2 问题二模型的建立与求解

5.2.1 随机森林模型对初始数据处理

随机森林是一种集成的分类回归算法，由多个决策树组成，基于多个决策树，可以将已知数据通过一定的规则划分为不同的类别。在第二问中，题目要求我们基于母亲的身体特征和心理特征，对婴儿的行为特征进行归类。已知婴儿的行为特征是离散型变量，有三种类别，恰好符合随机森林的特征，故我们在 MATLAB 中针对母亲身体特征和心理特征的八类数据，对于前 250 个样本数据进行随机森林测试组训练，再对后 129 个样本进行检测，最终得到准确率为 52%的随机森林模型，并成功根据最后 20 组婴儿母亲的数据对婴儿的行为特征进行预测，其中 2、7、14、16 至 19 组婴儿呈现出安静型的特征，剩余婴儿呈现出中等型的特征。通过对随机森林的进一步评估，绘制混淆矩阵，得到在中等型行为特征婴儿的判断上，随机森林模型准确率达到 60%，其中在安静型婴儿行为特征的判断上，随机森林模型的准确率达到 50%。值得指出的是，数据中矛盾型婴儿在数据占比中仅占 11.5%（45 人）。如果某类样本的数量较少，该类别的样本可能会被较少考虑，从而导致模型在训练过程中被误导影响准确性，这可能是在此基础上模型在预测婴儿行为特征时判断矛盾型婴儿数量较少的原因。为了解决此问题，接下来将讲述通过对数据进一步加工后再分析的随机森林模型。

5.2.2 优化后随机森林模型对数据处理

样本分布不均可能会导致随机森林模型更可能会犯错，所以要解决样本分布不均的问题。本文我们采用两种方法：欠采样和重采样。矛盾型、中等型、安全型在比例上呈现 45:225:120 的特征，故可以通过对中等型进行欠采样和对矛盾型进行重采样，使三项数据保持大致 1:1:1 的比例，删去一半的中等型婴儿的数据，并对矛盾型婴儿重复采样，利用 Excel 的随机数生成函数进行随机打乱以排除干扰，最终再次放入随机森林模型进行分析。最终得到的数据综合了三项类别，混淆矩阵显示在矛盾型准确率上达到 78.3%，在中等型上为 39.4%，在安静型上为 53.6%，总体上准确率达到 54%，相比于优化前随机森林模型有略微提升。优化后的随机森林模型在预测后最后 20 组婴儿上得到了与优化前熵值更高的答案[3,2,2,2,2,2,1,2,2,2,1,2,2,3,2,3,2,3,3,2]，即编号为 397、401 的婴儿呈现矛盾型，编号为 391、404、406、408、409 的婴儿呈现安静型，剩余均为中等型。模型优化后矛盾型、中等型，安静型比例呈现 1:7:2，整体上与优化前的数据一致并体现出了更高的熵值，故优化后的随机森林模型具有较高的可信度。但目前预测所得与已有的数据仍有较大差异，故使用多元逻辑回归模型再次进行计算。

5.2.3 主成分分析法降维

针对母亲身体指标和心理指标受多项变量影响的情况，我们选择主成分分析法筛选出起主导作用的变量进行降维，再建立多元逻辑回归模型，具体操作如下：

将母亲的母亲年龄、婚姻状况、教育程度、妊娠时间、分娩方式、CBTS、EPDS、HADS 记为 X1,X2,...X8，X=（X1,...X8）'是 8 维随机向量，X 为标准化后的矩阵。通过公式（2),获得协方差矩阵 R

$$R=\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\overline{x_i})^2\sum_{k=1}^{n}(x_{kj}-\overline{x_j})^2}} \tag{2}$$

用 MATLAB 计算特征值λ，从而分别由公式（3)、公式（4）获得贡献率及累计贡献率。

$$\text{贡献率}=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\ldots 8) \tag{3}$$

$$\text{累计贡献率}=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}(i=1,2,\ldots 8) \tag{4}$$

最终得到下表：

表五：主成分分析表

	a1	a2	a3	a4	a5	a6	a7	a8
母亲年龄	-0.0421	-0.0338	0.4580	-0.3633	0.5611	-0.3499	0.4555	0.1035
婚姻状况	0.0260	-0.0358	0.1964	-0.2969	-0.8143	-0.3514	0.2844	0.0627
教育程度	-0.0016	-0.0071	-0.4568	0.5513	0.0707	-0.6038	0.3184	0.1279
妊娠时间（周数）	0.0740	0.0105	0.5990	0.3888	-0.0229	-0.3813	-0.5726	0.1031
分娩方式	0.0429	-0.0281	0.4184	0.5617	-0.1170	0.4662	0.5241	0.0327
CBTS	-0.3991	-0.7164	-0.0430	-0.0132	-0.0172	0.1005	-0.0851	0.5548
EPDS	0.7980	0.0252	-0.0858	-0.0928	0.0355	0.0985	-0.0118	0.5792
HADS	-0.4406	0.6948	-0.0057	-0.0208	-0.0373	0.0834	-0.0173	0.5603
贡献性	0.3220	0.1580	0.1468	0.1213	0.0996	0.0925	0.0381	0.0217
累计贡献性	0.3220	0.4800	0.6268	0.7481	0.8477	0.9402	0.9783	1.0000

从上表可知，前五个的累计贡献率为 84.7%，第一主成分 F1 在除分娩方式和教育程度上的其他变量上都有较高的载荷；第二主成分在 CBTS 和 HADS 上有较高的载荷，其余载荷量近似。经过对前五个累计贡献率的分析比较，可达降维目的，进一步进行多元逻辑回归检验。

5.2.4 多元逻辑回归模型建立与结果预测

为得到婴儿行为特征与母亲的身体指标、心理指标的关系，需要建立多元逻辑回归模型。

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_6x_6 \tag{5}$$

其中母亲的母亲年龄、教育程度、妊娠时间、CBTS、EPDS、HADS 记为分别记为 x1，x2，...，x6。通过 SPSS 进行多元逻辑回归分析，得到准确率为 70.3%的多元 Logistic 回归模型，其中矛盾型预测准确性为 35.6%，中等型预测准确性为 87.6%，安静型预测准确性为 50.8%，用此对最后 20 组婴儿的行为特征进行预测，编号为 392、394、399 的婴儿呈现矛盾型，编号为 391、396、397、401、405、406、409 的婴儿呈现安静型，其余均为中间型，矛盾型、中间型、安静型比例呈现 3：10：7，整体上矛盾型、中间型、安静型比例约为 2：10：5，二者较为一致。

表六：问题二最后 20 组婴儿行为特征预测

编号	391	392	393	394	395	396	397	398	399	400
行为特征	安静型	矛盾型	中等型	矛盾型	中等型	安静型	安静型	中等型	矛盾型	中等型
编号	401	402	403	404	405	406	407	408	409	410
行为特征	安静型	中等型	中等型	中等型	安静型	安静型	中等型	中等型	安静型	中等型

5.3 问题三模型的建立与求解

5.3.1 中等型的最少治疗费用求解

基于 238 号的婴儿母亲的 CBTS、EPDS、HADS 三项指标分数分别为 15、22、18 分，婴儿的行为特征呈现矛盾型。故基于问题二的多元逻辑回归模型，我们得知在治疗过程中母亲的身体特征不会发生较大的改变，故通过穷举法列举三项指标从当时值降为 0 的过程中的每一次的治疗方案，一共有 16*23*19 种，即 6992 种方案。通过随机森林进行判断并将结果输出在 Excel 中，我们可以得到在分数降低过程中所有中等型时的所有分数指标，再基于原来 238 号婴儿的分数去计算所有中等型婴儿的治疗费用，经过排序，得到最低费用的解决方案。

经过模型的分析，我们发现在 6992 种方案中，能够将婴儿从矛盾型转变为中等型的方案一共有 662 种，转变为安静型的方案一共有 6324 种。在转变为中等型的 662 个方案中，价格最优为 3085 元，对应方案为 CBTS 15 分,EPDS 19 分,HADS18 分，即 CBTS 降 0 分，EPDS 降 3 分，HADS 降 0 分。

5.3.2 安静型后的治疗方案求解

基于随机森林判断模型，在 238 号婴儿从矛盾型转化为安静型的 6324 个方案中，我们将所有方案的价格进行排序，得到最优方案的价格为 6750.67 元，其对应方案为 CBTS 14 分，EPDS 22 分，HADS 16 分，即 CBTS 降低 1 分，EPDS 降低 0 分，HADS 降 2 分。

5.4 问题四模型的建立与求解

5.4.1 数据处理

首先对婴儿睡眠时间，睡醒次数，入睡方式建立评判标准。由附件中的补充说明分析可知：婴儿睡眠时间为极大型指标，睡醒次数为极小型指标，入睡方式从 1 至 5 显示出婴儿入睡的环境要求逐步减小，抗干扰能力增强，时间规律呈现显著趋势，这里我们偏向于解读为婴儿入睡方式的数值与其睡眠质量呈现正相关，因此我们将入睡方式从类别变量转换为数值变量且评判为极大型指标。

完成指标特点判定后，对睡醒次数指标进行正向化处理，使其由极小型指标转化

为极大型指标。完成对指标方向统一后，对这三个指标其进行 Min-Max 标准化处理，将所有数值收敛至 0-1 区间后，数据处理完成。

5. 4. 2 秩和比综合评价法 RSR 评价分类

秩和比综合评价法是一种多指标综合评价方法，用于将多个自变量指标的排名转化为秩和比值。我们首先将婴儿的睡眠数据（睡眠时间、睡醒次数、入睡方式）进行转化和标准化，对于连续性数据如睡眠时间，转化为无量纲的相对指标，对于入睡方式等离散型数据，可直接作为自变量也可转化为数值型数据，输入到多元回归模型中。最后根据各项数据的数值大小对评价对象进行排名求出秩（R），

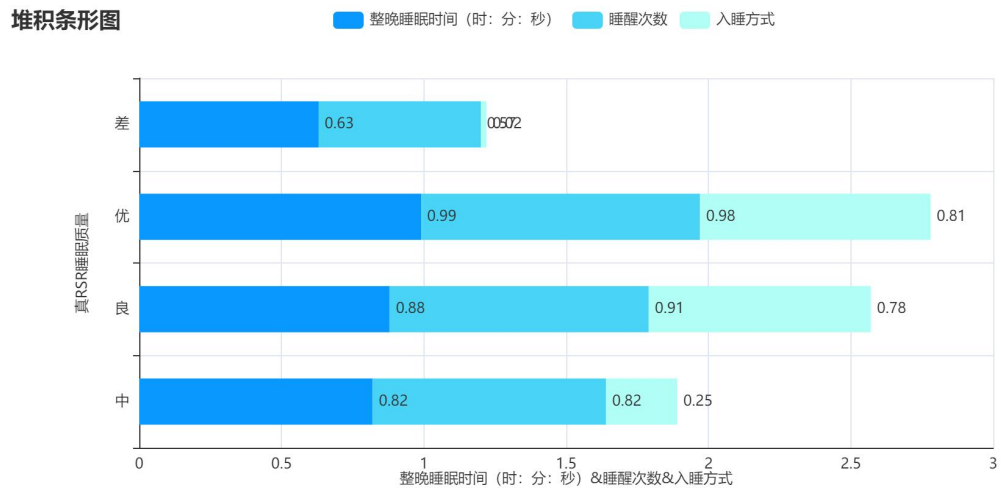
$$RSR = \frac{R1 + R2 + \ldots + Rn}{n \cdot \frac{(n+1)}{2}}$$

并根据上图公式计算得出指标的秩和比值（RSR），将 RSR 值转化为 probit 值，probit 值能够反映该指标对整个结果的影响程度，即睡眠数据的小项对于整体睡眠质量的影响程度。获得 probit 值后，可通过 SPSSPRO 计算得到回归方程 $RSR(WRSR) = -0.191 + 0.177 \times probit$，并按照回归方程计算的 $RSR / WRSR$ 估计值，将评价对象分为优、良、中、差四种类型。

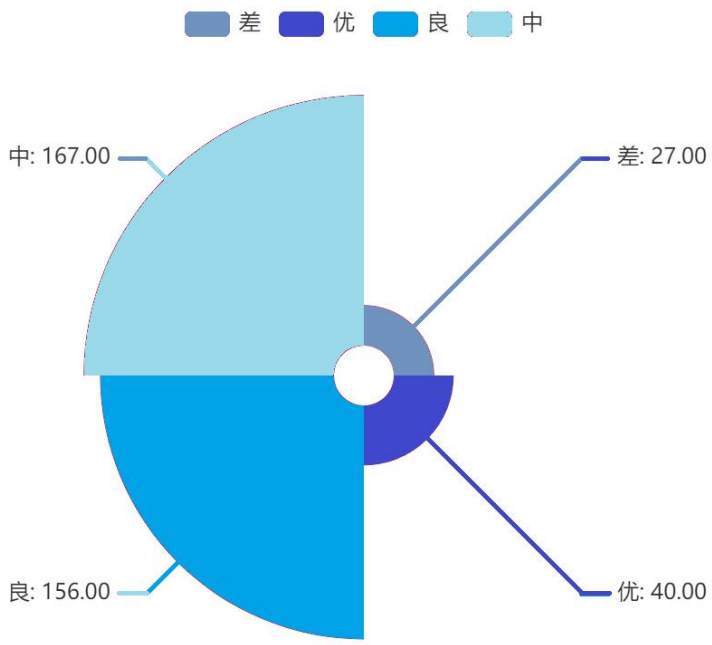
表七：分档排序临界值表格

睡眠质量	百分位临界值	Probit	RSR 临界值（拟合值）
差	<6.681	<3.5	<0.4282
中	6.681～	3.5～	0.4282～
良	50.000～	5～	0.6934～
优	93.319～	6.5～	0.9587～

图三：RSR 睡眠质量与各小项的关系图



图四：四个睡眠质量类型分布图



5. 4. 3 睡眠质量与母亲身体指标和心理指标的关联模型与结果预测

根据已获得的睡眠质量数据和对应母亲的身体和心理指标，将数据再次代入到多元逻辑回归模型中进行训练和测试，以睡眠质量为因变量，母亲身体和心理指标为自变量，建立得到准确率为 75.6%的多元逻辑回归模型，其中睡眠质量为差的预测准确率为 88.9%，睡眠质量为中的预测准确率为 79.6%，睡眠质量为良的预测准确率为 67.9%，睡眠质量为优的预测准确率为 80.0%，用此对最后 20 组婴儿的睡眠质量进行预测，其中编号为 394、396、404、407、409 的婴儿睡眠质量为差，编号为 391、393、395、398、399、410 的婴儿睡眠质量为中，编号为 392、397、400、402、403、405、408 的婴儿睡眠质量为良，编号为 401、406 的婴儿睡眠质量为优。

表八：问题四最后 20 组睡眠质量预测分类

编号	391	392	393	394	395	396	397	398	399	400
睡眠质量	中	良	中	差	中	差	良	中	中	良
编号	401	402	403	404	405	406	407	408	409	410
睡眠质量	优	良	良	差	良	优	差	良	差	中

5.5 问题五模型的建立与求解

5.5.1 穷举法数据逻辑再回归

问题三已经建立了 238 号婴儿母亲的心理治疗结果及付费情况的穷举数据表。将该数据中的中等型和安静型治疗方案筛选出来，重新代入婴儿睡眠质量关于母亲身体指标和心理指标的多元逻辑回归模型，在所得结果中筛选出婴儿睡眠质量为优的治疗方案。对已得到的心理指标治疗结果数据进行反向的标准化和正向化，使其还原为原始数据得分。最后代入问题三中得出的线性方程计算得出最终费用并从中进行排序和筛选，选取出价格最低的治疗方案。

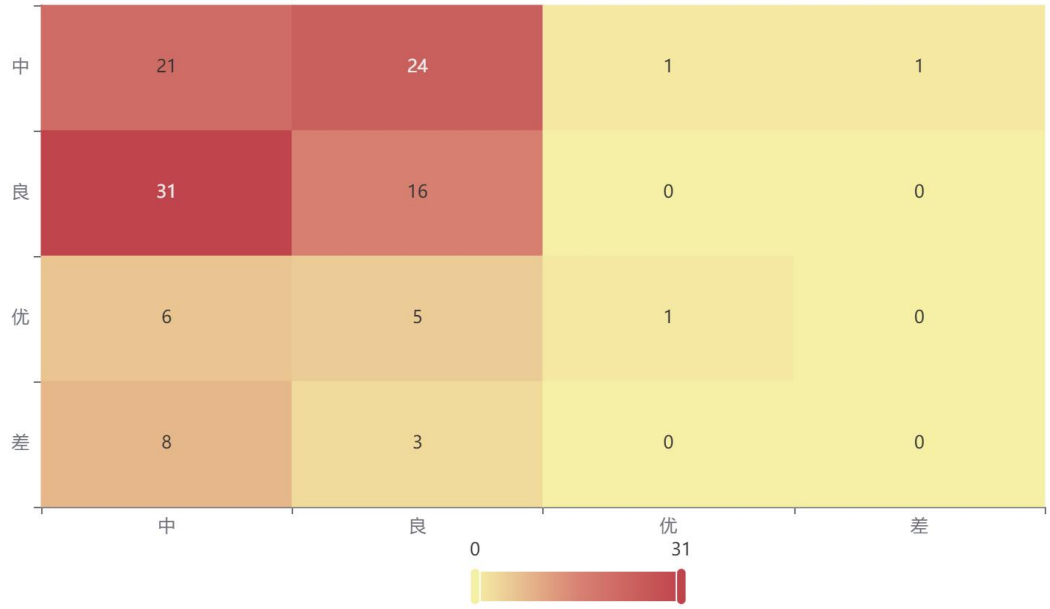
5.5.2 中等型—优治疗方案

经过随机森林模型计算，若在中等型婴儿的基础上要得到得到睡眠质量为优的最低价治疗方案，则价格为 6915 元，其中 CBTS 降 0 分，EPDS 降 5 分，HADS 降 1 分。

5.5.3 安静型—优治疗方案

若在安静型婴儿的基础上要得到睡眠质量为优的最低价治疗方案，则价格为 10046.34 元，其中 CBTS 降 2 分，EPDS 降 7 分，HADS 降 1 分。

图五：多元逻辑回归模型混淆矩阵



六、模型的分析与检验

6.1 灵敏度分析

为了评价模型的灵敏度，我们需要关注模型对自变量（母亲的身体特征和心理特征数据）变化的敏感程度。首先我们定义自变量的变化范围为[0,1]，并计算出原始逻辑回归模型的准确率为 39.55%，我们将每个自变量在设定变化范围内进行变化，并获取自变量变化后的模型预测结果。经观察可得，模型呈现出 41%的准确率，这表明自变量的变化会导致准确率的偏差，即我们的模型对数据的敏感性较强，当面对数据的偏差时，准确性会受较大影响。同时我们对灵敏度分析的灵敏度输出进行检测，正的灵敏度表示增加自变量会导致对模型的预测结果产生正面影响，而负的灵敏度表示减少自变量会导致预测结果产生负面影响。经过对八个自变量的灵敏度分析的结果可知，第 1、4、7、8 个自变量的灵敏度较大，说明他们会对模型的预测结果产生较大影响，第 3、6 个自变量的灵敏度为负，说明它们对模型的预测结果会产生负的影响，第 2 和 5 个自变量的灵敏度值为 0，说明他们对模型的预测结果没有明显的影响。下表表示各个自变量的灵敏度具体数值。数值的绝对值越大，说明该变量对整体模型应县越大。

表九：灵敏度检验表

变量类型	母亲年龄	婚姻状况	教育方式	妊娠时间	分娩方式	CBTS	EPDS	HADS
灵敏度	0.149	0	-0.075	0.223	0.075	-0.075	0.149	0.149

6.2 误差分析

误差分析指对模型预测结果与实际结果进行比对来判定模型的准确率。对多分类逻辑回归模型进行误差分析，在逻辑回归模型测试中，我们得到所有错误样本的汇总数据，其中误差类别为矛盾型的数量为 39，说明在这个类别上预测错误率较高，而类别为中等型和安静型的数量分别为 3 和 0，说明模型在这个类别上错误率较低，同时对于误差类别为矛盾型错误率较高的问题，之后将在模型的缺点和改进中进行具体阐述。

表十：误差分析表

行为特征	矛盾型	中等型	安静型
误差数	39	3	0

七、模型的评价

7.1 模型的优点

本文使用了随机森林、多元 Logistic 回归模型，秩和比综合评价法、穷举法数据逻辑再回归。根据各个模型和方法的表现与输出结果，我们总结出了以下各个模型的优点。

- 模型具有较高的可解释性，随机森林回归模型和多分类逻辑回归模型在测试集上分别呈现 60%和 73%的准确性，基本能够解释行为特征与母亲身体和心理

指标的关系，验证了模型的可靠性。

- 模型具有良好的泛化能力，随机森林模型和多分类逻辑回归模型在面对复杂的新的环境的数据，能够基于各项指标的特点和权重，给出分类评价，体现了模型能够广泛应用于其他情况。
- 模型训练基于合理的训练集合，在训练随机森林模型和多分类逻辑回归模型的过程中，考虑到了不同样本样本存在分布不均的问题，通过将数据进行欠采和重采的方式，得到了三类样本分布较均匀后的训练模型，此基础上可以减少由于分布不均到一个影响模型判断的准确率，使模型训练更有效。
- 模型具有较高的实际意义，考虑到了主要因素的影响。通过秩和比综合评价法来提取相关性较为显著的指标，在建立逻辑回归模型时更多考虑相关性更为显著的指标，使输出结果更贴近实际值，提高了模型的准确性。
- 逻辑回归模型基于穷举法提高了方案筛选效率，通过穷举所有的治疗方案及三项心理指标的分数情况，直接获得所有治疗方案以及对应价格，能够更广泛应用于其他情况，构建一个可供参考的图表。

7.2 模型的缺点

- 数据之间比例不平衡导致的多分类Logistic回归模型和随机森林模型有误差，尽管我们已通过重复采样和加强欠样来减少数据不平衡对分类模型的影响，但对模型准确率的提升不显著，故数据本身分布不均会影响最终的预测结果。
- 模型的准确率仍然具有误差，判断行为特征之间的分界线较为模糊。在治疗方案模型中，多分类逻辑回归对穷举法所有结果生成的行为特征具有一定误差，并不能完全拟合真实情况，可能会导致最终的治疗方案价格高于应有的价格。

7.3 模型的改进

- 将数据进行更复杂的预处理来提高模型准确率，比如将一组数据分为多组数据给随机森林模型和多元逻辑回归模型进行训练和测试，在各种方案中持续提升随机森林模型的准确率。
- 关于随机森林模型的改进，可以强化重要的特征，通过选择最相关的特征来提高准确率，在秩和比综合评价法分类的基础上，可以利用信息熵的实时反馈来判断随机森林分类的情况，以获取具有更大优势的特征信息。

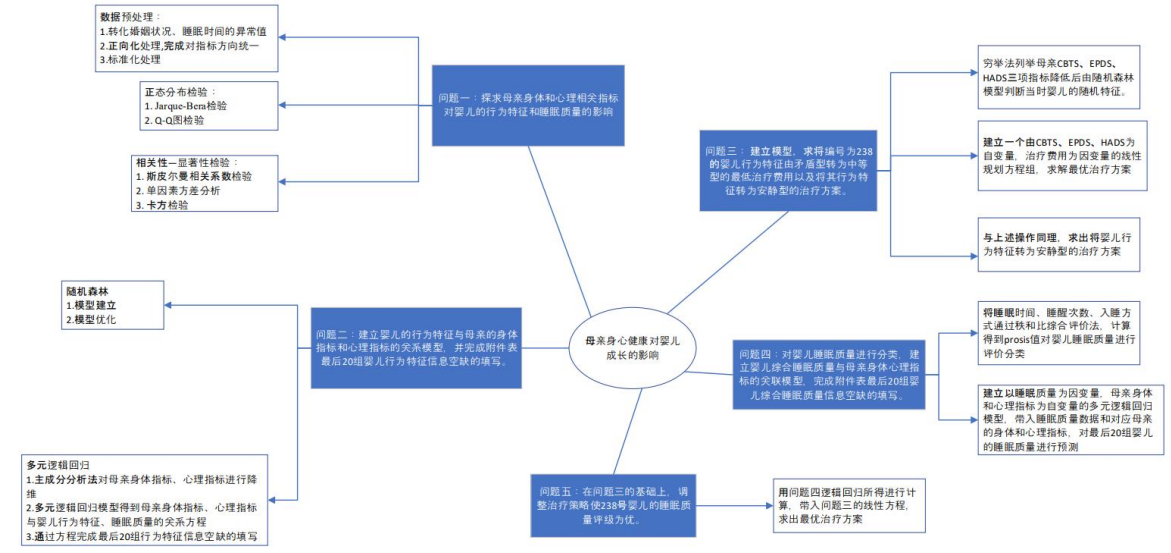
八、参考文献

[1] Winnicott, D. W. , Babies and Their Mothers，London, Free Association Books, 1988.
[2] 田凤调. 秩和比法及其应用[M]. 北京 中国统计出版社 1993.

附录

附录 1

思维导图



附录 2

Matlab 对数据预处理后的附件进行正向化和标准化

```
clear;clc
%导入数据
data = xlsread('附件.xlsx')%附件已经经过数据预处理，去除了异常值，删除了后二十组空白的数据并将字符串转化为逻辑变量，矛盾型为 1，中等型为 2，安静型为 3
[n,m] = size(data)
data_2 = data
%正向化处理，对于 EPDS、HADS、CBTS 三组数据，是极小型数据，需要进行正向化变为极大型数据。
Positivicolumn = [6,7,8];
for i = 1 : size(Positivicolumn,2)
data(:,Positivicolumn(i)) = Min2Max(data(:,Positivicolumn(i)));
end
disp(data)
%使用 Min-Max 标准化处理形成新矩阵，需要标准化的为连续性变量，即母亲年龄（1），妊娠时间（4），CBTS（6），EPDS（7），HADS(8)，整晚睡眠（11），睡醒次数（12）
Positivicolumnn = [1,4,6,7,8,11,12];
for i = 1 : size(Positivicolumnn,2)
data(:,Positivicolumnn(i)) = (data(:,Positivicolumnn(i)) - min(data(:,Positivicolumnn(i))))/(max(data(:,Positivicolumnn(i)))-min(data(:,Positivicolumnn(i))))
end
disp('为我们所得的标准化矩阵为')
```

```
disp(data)
%输出到新的 excel 表格
% 指定要保存的文件名和工作表名
filename = '附件 output.csv';
% 使用 xlswrite 函数将矩阵写入 Excel 可读取的 csv 文件
csvwrite('附件 output.csv', data);

function col = Min2Max(x)
col = max(x) - x;
end
```

附录 3

问题二中 matlab 回归预测逻辑模型代码及输出预测值

```
clc;clear
%多分类逻辑回归预测模型 matlab 代码部分
%使用 matlab 内置函数 mnrfit 和 mnrval
% 假设训练集数据为 X_train 和 y_train
% X_train 是训练集的自变量，即母亲的身体特征和心理特征数据
% y_train 是训练集的因变量，即婴儿的行为特征
%在进行训练时，发现婴儿三个行为特征的占比并不均衡，中等型占据大多数，
% 这会影响逻辑回归预测模型的准确性，
% 故我们通过重采样：通过欠采样（随机删除占比较多的类别样本和复制占比较少的类别样本）来平衡训练集中的类别比例。
data = xlsread('原数据.xlsx')
X_train = data(1:100, 1: 8);
y_train = data(1:100,14);
% 使用 mnrfit 内嵌函数训练多分类逻辑回归模型
model = mnrfit(X_train, y_train);
X_test = data(101:end,1: 8)
y_test = data(101:end,14) %X_test 与 y_test 是训练集，即 200 个测试集
% 输出预测测试集的类别
predicted_test_labels = mnrval(model, X_test);
[~, predicted_test_labels] = max(predicted_test_labels, [], 2);
% 输出测试集的类别与实际类别比较，得到准确率
accuracy = sum(predicted_test_labels == y_test) / numel(y_test);
%导入缺失的二十组婴儿数据
res1 = xlsread('输出文件.xlsx')
X_baby = res1(1:end, 1: 8)
% 预测挖空数据的类别
```

```
    predicted_missing_labels = mnrval(model, X_baby);
    [~, predicted_missing_labels] = max(predicted_missing_labels, [], 2);
    %结果准确率为 53%
    %输出二十组婴儿行为特征为：[2,3,2,2,3,2,3,2,2,3,2,2,2,2,3,2,3,2,2,2]
    %1 为矛盾型，2 为中等型，3 为安静型
```

附录 4

问题二中 matlab 随机森林模型代码及输出预测值

```
    clc;clear
    % train 是训练集的自变量，即母亲的身体特征和心理特征数据
    % train_result 是训练集的因变量，即婴儿的行为特征
    % test 是测试集的自变量，即母亲的身体特征和心理特征数据
    % train_result 是测试集的因变量，即婴儿的行为特征，1 为矛盾，2 为中等，3 为安静
    % 导入数据训练集和测试集的数据
    res = xlsread('原数据.xlsx');
    train = res(1:250, 1:8);%取前二百五十行的数据,行为样本数据，列为指标
    train_result = res(1:250, 14);%第十四列为婴儿的行为特征
    test = res(251:end, 1:8);%取二百五十行后的数据,行为样本数据，列为测试集的指标
    test_result = res(251:end, 14);%第十四列为将要对比的婴儿的行为特征答案
    % 使用 TreeBagger 内嵌函数建立随机森林模型，决策树为 100 棵
    numTrees = 100; % 随机森林中树的数量
    model = TreeBagger(numTrees, train, train_result, 'Method', 'classification');
    % 输出预测训练集的类别与 train_result 对比
    predicted_train_labels = predict(model, train);
    predicted_train_labels = str2double(predicted_train_labels);%将类别输出
    % 输出训练结果的准确率并进行输出
    train_accuracy = sum(predicted_train_labels == train_result) /
numel(train_result);
    disp(['训练集准确率：', num2str(train_accuracy)]);
    % 输出预测训练集的类别与 test_result 对比
    predicted_test_labels = predict(model, test);
    predicted_test_labels = str2double(predicted_test_labels);
    % 输出测试结果的准确率并进行输出
    test_accuracy = sum(predicted_test_labels == test_result) /
numel(test_result);
    disp(['测试集准确率：', num2str(test_accuracy)]);
    %导入缺失的二十组婴儿数据
    res1 = xlsread('输出文件.xlsx');
    X_baby = res1(:, 1:8);
```

```
% 预测二十组婴儿的行为特征
predict_baby = predict(model, X_baby);
predict_baby = str2double(predict_baby);
% 输出预测的婴儿的类别：1 or 2 or 3
disp('预测的类别标签：');
disp(predict_baby);
%结果准确率为 56%
%输出二十组婴儿行为特征为：[2,3,2,2,2,2,3,2,2,2,2,2,2,2,3,2,3，3,3,3,2]
%1 为矛盾型，2 为中等型，3 为安静型
% 计算混淆矩阵
C = confusionmat(test_result, predicted_test_labels);
% 输出混淆矩阵
disp('混淆矩阵：');
disp(C);
%混淆矩阵将以矩阵的形式显示，其中行表示真实类别，列表示预测类别。
% 每个元素表示预测为某个类别的样本在真实类别中的数量。
,2,3,3,2,2，3,2,3,2]
%1 为矛盾型，2 为中等型，3 为安静型

%混淆矩阵结果如下
% 输出混淆矩阵
disp('混淆矩阵：');
disp(C);
混淆矩阵：
     0    17     6
     2    87    21
     3    39    15
1 的准确率为 0%，2 的准确率为 60.8%，3 的准确率为 50%
```

附录 5

问题三：穷举法数值多元逻辑回归

```
clear;clc;
%多分类逻辑回归预测模型 matlab 代码部分
%使用 matlab 内置函数 mnrfit 和 mnrval
% 假设训练集数据为 X_train 和 y_train
% X_train 是训练集的自变量，即母亲的身体特征和心理特征数据
% y_train 是训练集的因变量，即婴儿的行为特征
% X_test 是测试集的自变量，即母亲的身体特征和心理特征数据
% y_test 是测试集的因变量，即婴儿的行为特征，1 为矛盾，2 为中等，3 为安静
%在进行训练时，发现婴儿三个行为特征的占比并不均衡，中等型占据大多数，
% 这会影响逻辑回归预测模型的准确性，
```

```
    % 故我们可以通过重采样和欠采样（随机删除占比较多的类别样本和复制占比较少的类别样本）
来平衡训练集中的类别比例。
    data = xlsread('原数据_1.xlsx')
    X_train = data(1:200, 1: 8);
    y_train = data(1:200,14);
    % 使用 mnrfit 内嵌函数训练多分类逻辑回归模型
    model = mnrfit(X_train, y_train);
    X_test = data(201:end,1: 8)
    y_test = data(201:end,14) %X_test 与 y_test 是训练集，即 200 个测试集
    % 输出预测测试集的类别
    predicted_test_labels = mnrval(model, X_test);
    [~, predicted_test_labels] = max(predicted_test_labels, [], 2);
    % 输出测试集的类别与实际类别比较，得到准确率
    accuracy = sum(predicted_test_labels == y_test) / numel(y_test);
    %导入待测情况文件
    res1 = xlsread('输出文件.xlsx')
    X_baby = res1(1:end, 1: 8);
    f=@(a)870.67*a+200;
    g=@(b)695*b+500;
    h=@(c)2440*c+300;
    % 初始化数据
    data = [15, 22, 18];
    alldata = [];
    % 穷举法:全部治疗情况
    for i = data(1):-1:0
    for j = data(2):-1:0
    for k = data(3):-1:0
    alldata = [alldata; i, j, k];
    end
    end
    end
    %alldata 为全部治疗情况
    %%
    for s=1:1:6992
    %对 alldata 数据进行正向化，标准化处理，否则无法与逻辑回归算法的数据匹配
    alldata(s,1)=(21-alldata(s,1))/21;
    alldata(s,2)=(28-alldata(s,2))/28;
    alldata(s,3)=(20-alldata(s,3))/20;
    end
    result=[];
    p=0;
    for t=1:1:6992
    %建立 for loop 调用 alldata 处理过的数据进行预测
    X_baby(1,6:8)=alldata(t,1:3)
```

```
    % 预测该情况下的行为特征
    predicted_missing_labels = mnrval(model, X_baby);
    [~, predicted_missing_labels] = max(predicted_missing_labels, [], 2);
    if predicted_missing_labels(1,1)==2 %若判断结果为 2（中等型）则自动归入新矩阵并记录其治疗后得分和费用
    p=p+1;
    result(p,1:5)=[2,21-21*alldata(t,1),28-28*alldata(t,2),20-20*alldata(t,3),f(-6+21*alldata(t,1))+g(-6+28*alldata(t,2))+h(-2+20*alldata(t,3))];
    elseif predicted_missing_labels(1,1 )==3
    p=p+1;
    result(p,1:5)=[3,21-21*alldata(t,1),28-28*alldata(t,2),20-20*alldata(t,3),f(-6+21*alldata(t,1))+g(-6+28*alldata(t,2))+h(-2+20*alldata(t,3))];
    else
    end
    end
    T = array2table(result)
```

附录 6

问题三中判断心理指标分数降低后婴儿行为特征是否降为中等型或安静型随机森林模型代码

```
    clc;clear
    % train 是训练集的自变量，即母亲的身体特征和心理特征数据
    % train_result 是训练集的因变量，即婴儿的行为特征
    % test 是测试集的自变量，即母亲的身体特征和心理特征数据
    % train_result 是测试集的因变量，即婴儿的行为特征，1 为矛盾，2 为中等，3 为安静
    % 导入数据训练集和测试集的数据
    res = xlsread('原数据_1.xlsx');%
    train = res(1:250, 1:8);%取前二百五十行的数据,行为样本数据，列为指标
    train_result = res(1:250, 14);%第十四列为婴儿的行为特征
    test = res(251:end, 1:8);%取二百五十行后的数据,行为样本数据，列为测试集的指标
    test_result = res(251:end, 14);%第十四列为将要对比的婴儿的行为特征答案
    % 使用 TreeBagger 内嵌函数建立随机森林模型，决策树为 50 棵
    numTrees = 100; % 随机森林中树的数量
    model = TreeBagger(numTrees, train, train_result, 'Method', 'classification');
    % 输出预测训练集的类别与 train_result 对比
    predicted_train_labels = predict(model, train);
    predicted_train_labels = str2double(predicted_train_labels);%将类别输出
```

```
    % 输出训练结果的准确率并进行输出
    train_accuracy = sum(predicted_train_labels == train_result) /
numel(train_result);
    disp(['训练集准确率：', num2str(train_accuracy)]);
    % 输出预测训练集的类别与 test_result 对比
    predicted_test_labels = predict(model, test);
    predicted_test_labels = str2double(predicted_test_labels);
    % 输出测试结果的准确率并进行输出
    test_accuracy = sum(predicted_test_labels == test_result) /
numel(test_result);
    disp(['测试集准确率：', num2str(test_accuracy)]);
    %导入缺失的二十组婴儿数据
    res1 = xlsread('分界线检验.xlsx');
    X_baby = res1(:, 1:8);
    % 预测二十组婴儿的行为特征
    predict_baby = predict(model, X_baby);
    predict_baby = str2double(predict_baby);
    % 输出预测的婴儿的类别：1 or 2 or 3
    disp('预测的类别标签：');
    disp(predict_baby);
    %结果准确率为 56%
    %输出二十组婴儿行为特征为：[2,3,2,2,2,2,3,2,2,2,2,2,2,2,3,2,3，3,3,3,2]
    %1 为矛盾型，2 为中等型，3 为安静型
    % 计算混淆矩阵
    C = confusionmat(test_result, predicted_test_labels);
    % 输出混淆矩阵
    disp('混淆矩阵：');
    disp(C);
    %淆矩阵将以矩阵的形式显示，其中行表示真实类别，列表示预测类别。
    % 每个元素表示预测为某个类别的样本在真实类别中的数量。
```

附录 7

问题四中睡眠质量与母亲身体和心理指标随机森林模型代码

```
    %本代码用于计算在儿童睡眠质量为优时所需的治疗方案
    %首先需要通过多元逻辑回归模型去检测儿童睡眠质量与母亲身体和心理特征
    %然后我们用得到的回归模型去预测 20 组婴儿的睡眠质量
    %
    clc;clear
    % train 是训练集的自变量，即母亲的身体特征和心理特征数据
```

```
    % train_result 是训练集的因变量，即婴儿的睡眠质量
    % test 是测试集的自变量，即母亲的身体特征和心理特征数据
    % train_result 是测试集的因变量，即婴儿的行为特征，1 为矛盾，2 为中等，3 为安静
    % 导入数据训练集和测试集的数据
    res = xlsread('睡眠质量与身体心理指标.xlsx');
    train = res(1:250, 1:8);%取前二百五十行的数据,行为样本数据，列为指标
    train_result = res(1:250, 14);%第十四列为婴儿的睡眠质量
    test = res(251:end, 1:8);%取二百五十行后的数据,行为样本数据，列为测试集的指标
    test_result = res(251:end, 14);%第十四列为将要对比的婴儿的行为特征答案
    % 使用 TreeBagger 内嵌函数建立随机森林模型，决策树为 100 棵
    numTrees = 100; % 随机森林中树的数量
    model = TreeBagger(numTrees, train, train_result, 'Method',
'classification');
    % 输出预测训练集的类别与 train_result 对比
    predicted_train_labels = predict(model, train);
    predicted_train_labels = str2double(predicted_train_labels);%将类别输出
    % 输出训练结果的准确率并进行输出
    train_accuracy = sum(predicted_train_labels == train_result) /
numel(train_result);
    disp(['训练集准确率：', num2str(train_accuracy)]);
    % 输出预测训练集的类别与 test_result 对比
    predicted_test_labels = predict(model, test);
    predicted_test_labels = str2double(predicted_test_labels);
    % 输出测试结果的准确率并进行输出
    test_accuracy = sum(predicted_test_labels == test_result) /
numel(test_result);
    disp(['测试集准确率：', num2str(test_accuracy)]);
    %导入缺失的二十组婴儿数据
    res1 = xlsread('输出文件.xlsx');
    X_baby = res1(:, 1:8);
    % 预测二十组婴儿的睡眠质量
    predict_baby = predict(model, X_baby);
    predict_baby = str2double(predict_baby);
    % 输出预测的婴儿的类别：1 or 2 or 3 or 4
    disp('预测的类别标签：');
    disp(predict_baby);
    % 计算混淆矩阵
    C = confusionmat(test_result, predicted_test_labels);
    % 输出混淆矩阵
    disp('混淆矩阵：');
    disp(C);
    %淆矩阵将以矩阵的形式显示，其中行表示真实类别，列表示预测类别。
    % 每个元素表示预测为某个类别的样本在真实类别中的数量。
```

附录 8

问题五中对穷举出的母亲心理数据预测评判睡眠质量的随机森林模型

```
clc;clear
% train 为训练集的自变量，即婴儿的睡眠数据
% train_result 是训练集的因变量，即婴儿的睡眠质量，睡眠质量的数据来源于秩和比综合评价体系划分所
% test 是测试集的自变量，即母亲的心理特征数据
% train_result 是测试集的因变量，即婴儿的睡眠质量，1 为差，2 为中，3 为良，4 为优
% 导入数据训练集和测试集的数据
res = xlsread('睡眠质量与身体心理指标.xlsx');%
train = res(1:250, 1:8);%取前二百五十行的数据,行为样本数据，列为指标
train_result = res(1:250, 14);%第十四列为婴儿的睡眠质量
test = res(251:end, 1:8);%取二百五十行后的数据,行为样本数据，列为测试集的指标
test_result = res(251:end, 14);%第十四列为将要对比的婴儿的睡眠质量
% 使用 TreeBagger 内嵌函数建立随机森林模型，决策树为 100 棵
numTrees = 100; % 随机森林中树的数量
model = TreeBagger(numTrees, train, train_result, 'Method', 'classification');
% 输出预测训练集的类别与 train_result 对比
predicted_train_labels = predict(model, train);
predicted_train_labels = str2double(predicted_train_labels);%将类别输出
% 输出训练结果的准确率并进行输出
train_accuracy = sum(predicted_train_labels == train_result) / numel(train_result);
disp(['训练集准确率：', num2str(train_accuracy)]);
% 输出预测训练集的类别与 test_result 对比
predicted_test_labels = predict(model, test);
predicted_test_labels = str2double(predicted_test_labels);
% 输出测试结果的准确率并进行输出
test_accuracy = sum(predicted_test_labels == test_result) / numel(test_result);
disp(['测试集准确率：', num2str(test_accuracy)]);
%导入缺失的二十组婴儿数据
res1 = xlsread('分界线检验.xlsx');
X_baby = res1(:, 1:8);
% 预测二十组婴儿的行为特征
predict_baby = predict(model, X_baby);
predict_baby = str2double(predict_baby);
% 输出预测的婴儿的类别：1 or 2 or 3
disp('预测的类别标签：');
```

```
disp(predict_baby);
%结果准确率为 56%
%输出二十组婴儿行为特征为：[2,3,2,2,2,2,3,2,2,2,2,2,2,2,3,2,3，3,3,3,2]
%1 为矛盾型，2 为中等型，3 为安静型
% 计算混淆矩阵
C = confusionmat(test_result, predicted_test_labels);
% 输出混淆矩阵
disp('混淆矩阵：');
disp(C);
%淆矩阵将以矩阵的形式显示，其中行表示真实类别，列表示预测类别。
% 每个元素表示预测为某个类别的样本在真实类别中的数量。
```