所属类别	**2024** 年“华数杯”全国大学生数学建模竞赛	参赛编号
本科组		CM2400947

全国大学生数学建模竞赛论文模板

摘要

母亲的身心健康对婴儿的早期成长和发展具有重要影响。已有研究表明，母亲的身体状况以及心理健康（如压力、抑郁、焦虑等）不仅关系到婴儿的生理健康，还会影响婴儿的认知、情感、社会行为等多方面成长。婴儿的行为特征和睡眠质量常常作为衡量婴儿发展状况的重要指标。基于这方面的专业背景，本文对相关影响机制进行了研究与建模。

对于问题一, 本文通过分析 390 名婴儿及其母亲的身体和心理相关指标数据，研究了母亲因素对婴儿行为特征和睡眠质量的影响关系。

对于问题二, 基于婴儿行为问卷，将婴儿行为特征划分为安静型、中等型和矛盾型，建立了行为特征与母亲身体、心理指标之间的数学关系模型，并利用该模型预测了 20 组缺失婴儿行为类型。

对于问题三, 建立了 CBTS、EPDS、HADS 评分与治疗费用的关联模型，针对编号 238 的矛盾型婴儿，分析通过最小治疗费用使其行为特征由矛盾型转变为中等型和安静型的策略及调整路径。

对于问题四, 选取婴儿睡眠的多项指标，综合评判睡眠质量，并建立母亲各项指标与婴儿综合睡眠质量的关联模型，进而预测缺失数据婴儿的睡眠类别。

最后，结合模型结果，对如何提升 238 号婴儿的睡眠质量等级为优提出了具体的治疗方案和优化建议。

关键字： 母亲心理健康　婴儿成长　行为特征　睡眠质量　数学建模

一、 问题重述

1.1 问题背景

问题背景

1.2 问题要求

问题 1

问题 2

问题 3

问题 4

二、 问题分析

2.1 问题一分析

对于问题一，

2.2 问题二分析

对于问题二，

2.3 问题三分析

对于问题三，

2.4 问题四分析

对于问题四，

三、 模型假设

为简化问题，本文做出以下假设：

- 假设 1
- 假设 2
- 假设 3

四、 符号说明

符号	说明	单位
m	质量	kg
V	体积	m^3

五、 问题一的模型的建立和求解

5.1 数据预处理与特征工程

高质量的数据是构建有效模型的基础。针对原始数据集，我们进行了以下关键的数据预处理和特征工程步骤，旨在确保数据的规范性、一致性和可用性，为后续的模型构建奠定坚实基础：

5.1.1 睡眠时间格式转换

原始数据中婴儿“整晚睡眠时间”以“HH:MM”字符串格式记录，且存在“99:99”等异常值。我们将其统一转换为浮点数表示的小时数，并对异常值进行了妥善处理，通常将其视为缺失值或根据业务逻辑进行填充。

$$\text{时间}_{\text{小时}} = \text{小时} + \frac{\text{分钟}}{60}$$

此转换确保了睡眠时间数据的数值化和可计算性，使其能够被回归模型有效利用。

5.1.2 分类变量编码与映射

数据集中包含婚姻状况、教育程度、分娩方式、婴儿性别、入睡方式和婴儿行为特征等分类变量。为使其能够被回归模型处理，我们进行了以下操作：我们将原始的数值编码（如 1、2、3 等）映射为更具可读性的类别标签（如“未婚”、“已婚”；“哄睡法”、“自主入睡”）。对于无序的分类变量，如入睡方式、教育程度等，我们采用了独热编码将其转换为二元（0/1）指示变量。这有效避免了将分类变量误解释为有序数值，并消除了因任意数值赋值而引入的潜在偏差。例如，教育程度的“研究生”类别将被转换为一个独立的二元特征。值得一提的是，婴儿行为特征作为因变量，被映射为数值（如安静型 =0，中等型 =1，矛盾型 =2），以适应多项逻辑回归模型的输入要求。

5.1.3 数值型特征标准化

对于母亲年龄、妊娠时间、CBTS、EPDS、HADS 等数值型预测变量，我们均采用了标准差标准化（Z-score normalization）处理。

$$X' = \frac{X - \mu}{\sigma}$$

其中，X 是原始特征值，μ 是特征的均值，σ 是特征的标准差。标准化后的特征均值为 0，标准差为 1。此步骤至关重要，它有助于消除不同特征因量纲和取值范围差异带来的影响，确保模型优化过程中所有特征得到同等重视；加速模型收敛，尤其对于基于梯度下降的优化算法（如逻辑回归）；提高模型的可解释性和稳定性，避免某些特征因数值过大而主导模型。这些预处理步骤共同确保了数据的规范性、一致性和可用性，为后续的模型构建奠定了坚实基础。

5.2 探索性数据分析 (EDA) 结果

在正式构建模型之前，我们进行了全面的探索性数据分析，旨在揭示数据内部的结构、变量分布及其潜在关联，为后续模型选择与构建提供数据驱动的依据。

5.2.1 变量分布特征

我们首先对关键变量的分布特性进行了细致的探究，以全面了解数据的内在结构。对母亲的心理指标（CBTS、EPDS 和 HADS 得分）进行分析，如图 1 所示，这些指标的分布均呈现出明显的右偏态势。这清晰地表明，虽然研究样本中大部分母亲的心理健康状况总体良好，但仍有一部分母亲的得分相对较高，提示着她们可能存在不同程度的心理健康风险，需要进一步关注。

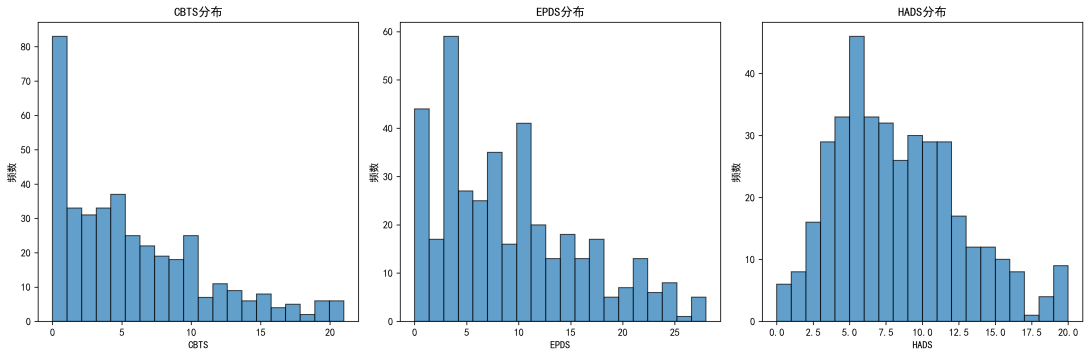

图 **1**　母亲心理指标分布

关于婴儿的睡眠质量，如图 2 所示，我们观察到婴儿的整晚睡眠时间主要集中在 10 至 12 小时这一区间，这与该年龄段婴儿正常的生理睡眠需求是基本吻合的。然而，在

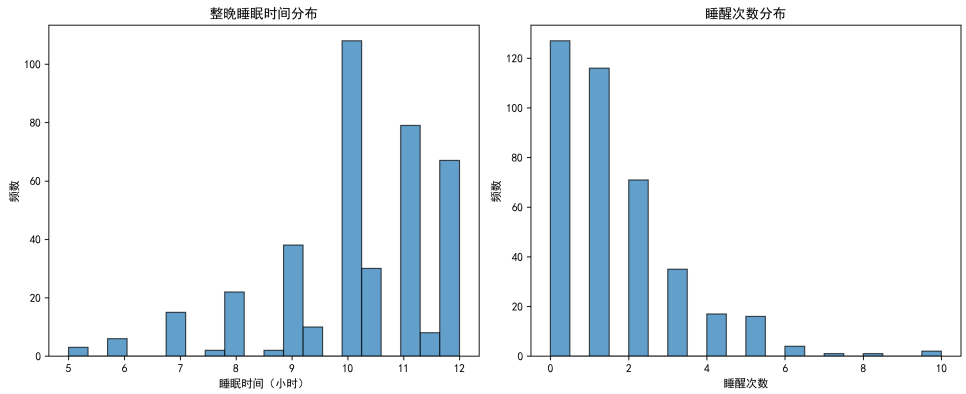

图 **2**　睡眠质量指标分布

夜间睡醒次数方面，虽然大多数婴儿的醒来次数维持在 0 至 2 次，但数据中也包含少数夜间醒来频率较高的婴儿，这可能暗示了其睡眠模式存在潜在的改善空间。

如图 3 所示，婴儿行为特征的分布呈现出显著的不均衡性。其中，“中等型”婴儿在样本中占据了最大比例（57.7%），其次是“安静型”婴儿（30.8%），而“矛盾型”婴儿的比例最低，仅为 11.5%。这种类别分布的明显差异，在后续构建模型时需要予以特别考量，以确保模型不会因数据不平衡而产生偏差。

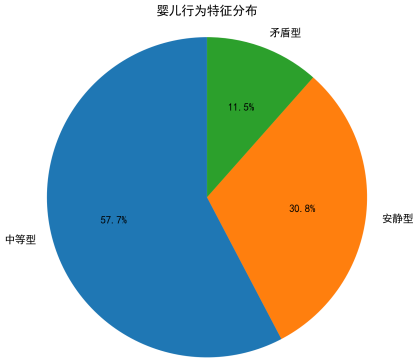

图 **3**　婴儿行为特征分布

5.2.2 相关性分析

为了洞察各数值型变量之间的相互关系，我们计算了皮尔逊相关系数，并以热力图形式直观展示了其关联强度。我们发现母亲的 EPDS 与 HADS 得分之间存在高达 0.79 的强相关性，CBTS 与 EPDS 为 0.78，而 CBTS 与 HADS 也达到 0.69。这种高度的正相关性强烈表明，抑郁、焦虑以及创伤后应激障碍等心理困扰并非孤立存在，而是倾向于同时出现，这为临床干预策略的制定提供了重要启示，即应采取综合性的评估与治疗方法。分析进一步揭示，如图4所示，母亲的各项心理指标（CBTS、EPDS、HADS）与婴儿的整晚睡眠时间呈现弱到中等程度的负相关，而与夜间睡醒次数则呈正相关。这一发

现初步勾勒出母亲心理压力水平越高，婴儿睡眠质量可能越差的变化规律，突显了母亲心理健康对婴儿健康发展的重要影响。

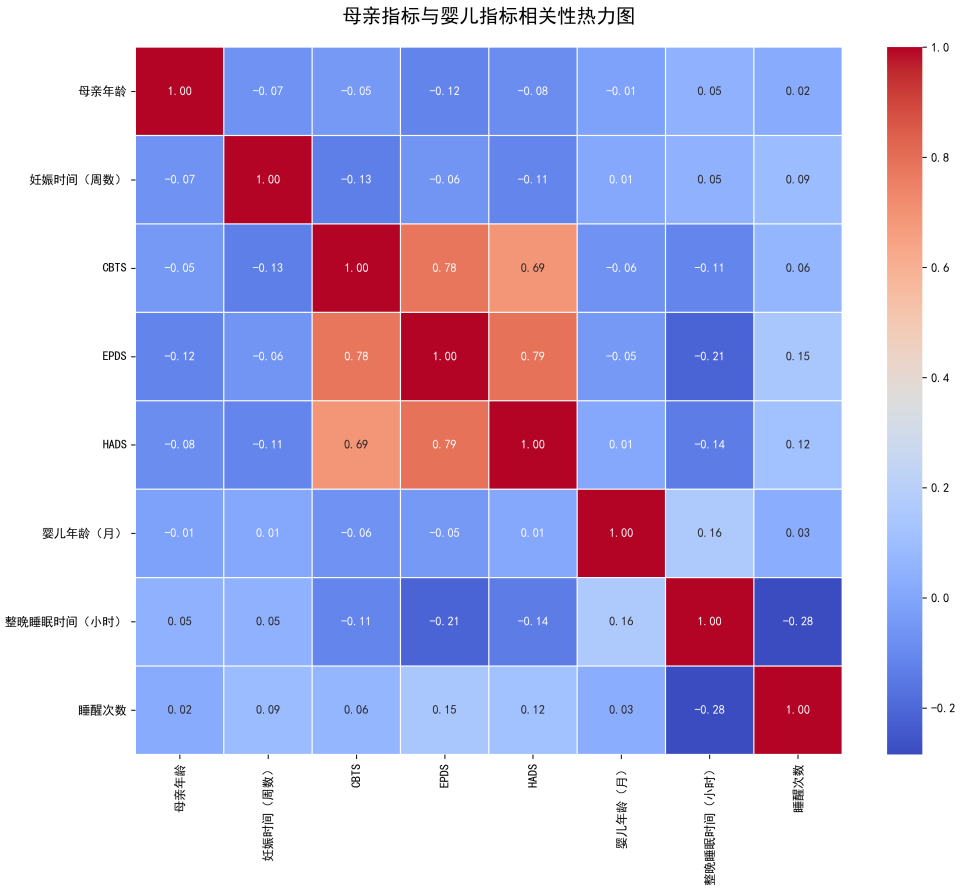

图 **4**　母亲指标与婴儿指标相关性热力图

5.2.3 分类变量对连续变量的影响 (ANOVA)

为了评估分类变量对婴儿睡眠质量指标（睡醒次数和整晚睡眠时间）的潜在影响，我们运用了方差分析（ANOVA），以检验不同类别之间是否存在显著差异。方差分析结果强烈支持入睡方式对婴儿睡眠质量指标具有显著影响。具体而言，入睡方式对婴儿夜间睡醒次数产生了显著影响（$F = 16.87, p < 0.001$）。如图 5 所示，能够自主入睡的婴儿，其夜间睡醒次数显著少于其他入睡方式的婴儿，这有力地表明自主入睡有助于提升婴儿夜间睡眠的连续性。

同样，入睡方式对婴儿的整晚睡眠时间也表现出显著影响（$F = 12.18, p < 0.001$）。如图 6 所示，自主入睡的婴儿普遍拥有更长的整晚睡眠时间，进一步印证了自主入睡对提升婴儿整体睡眠质量的积极作用。

值得注意的是，尽管我们考察了婚姻状况、教育程度、分娩方式以及婴儿性别等其他分类变量，但这些因素对婴儿睡眠质量的各项指标均未显示出统计学上的显著影响

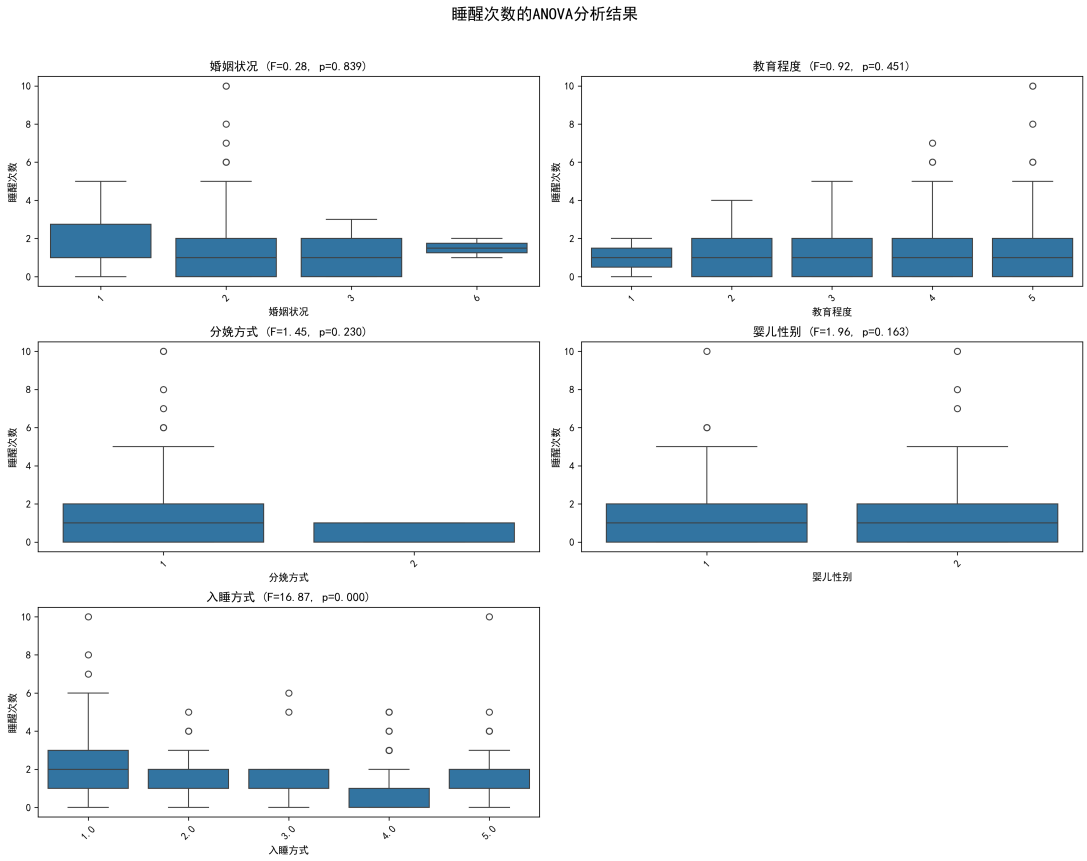

图 **5**　睡醒次数的 **ANOVA** 分析结果

（所有 $p > 0.05$）。这可能意味着，在影响婴儿睡眠的众多因素中，这些人口学和分娩相关变量的直接作用相对较弱。

5.2.4 分类变量间的关联性 (Chi-square)

我们进一步采用卡方检验来探究各个分类变量之间的关联性，以识别组别间是否存在非随机的分布模式。如图 7 所示，我们特别展示了那些具有统计显著性（$p < 0.05$）的分类变量对。卡方检验结果显示，母亲的婚姻状况与教育程度之间存在统计学上的显著关联（$p = 0.020$）。这提示在我们的样本中，已婚母亲的教育程度普遍高于未婚母亲，这或许反映了某些社会人口学背景下的普遍趋势。

更为重要的发现是，入睡方式与婴儿行为特征之间存在显著关联（$p = 0.014$）。具体分析表明，采用自主入睡方式的婴儿，其行为模式更倾向于表现为“安静型”特征。图 8 则以热力图的形式呈现了所有分类变量对的 p 值，p 值越小意味着关联性越强，进一步直观展示了这种关联。这一结果不仅强化了良好入睡习惯的重要性，也进一步揭示了其对婴儿整体行为发展所产生的积极影响。大多数分类变量之间未检测到统计学上的显著关联，这暗示了这些因素在很大程度上是相互独立的，从而简化了后续模型中的变量选择和解释。

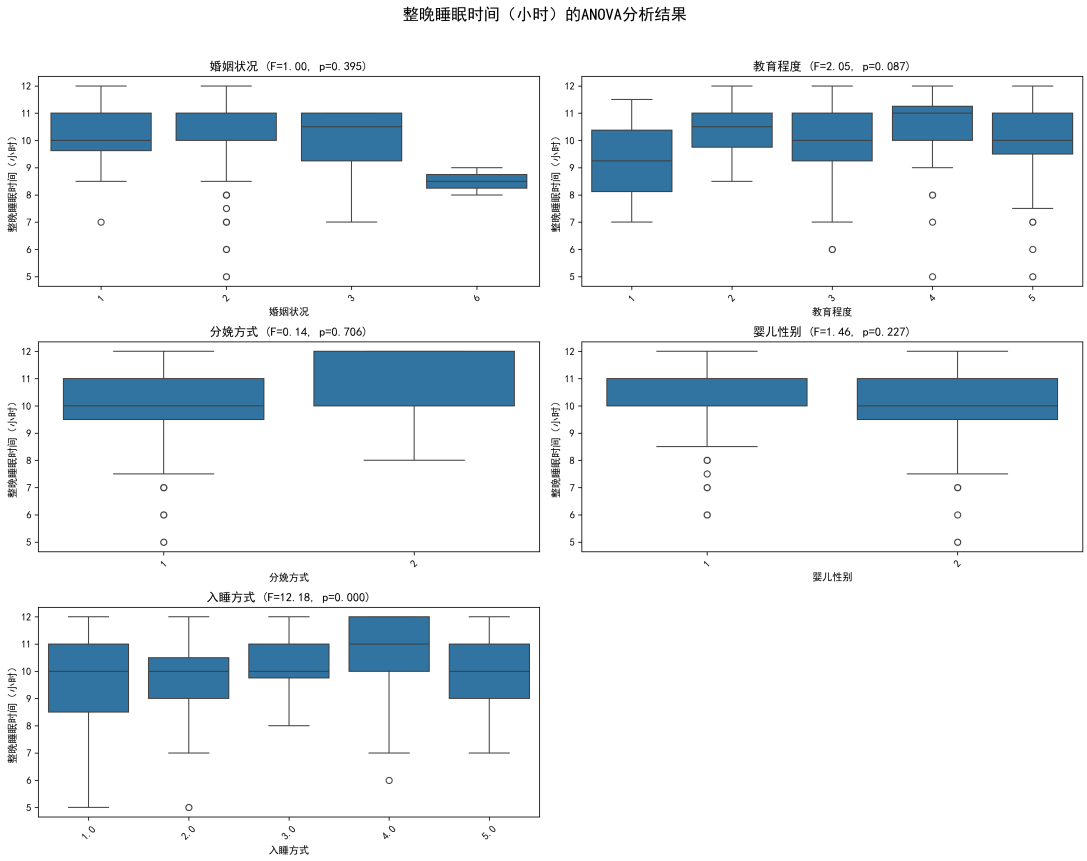

图 **6**　整晚睡眠时间（小时）的 **ANOVA** 分析结果

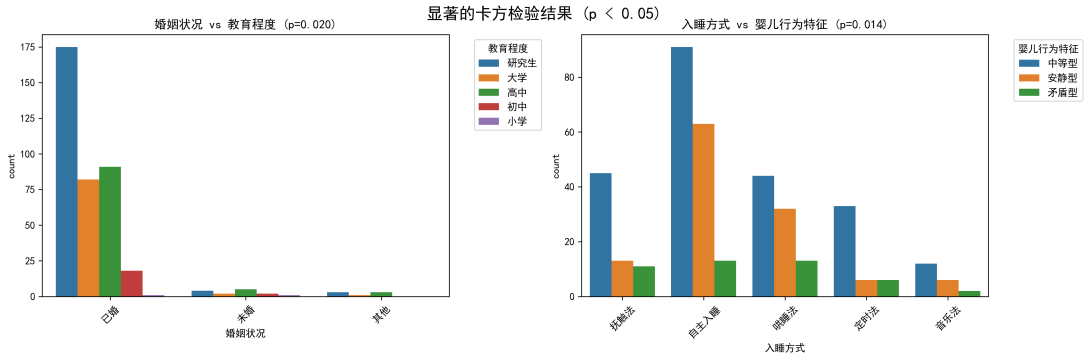

图 **7**　显著的卡方检验结果 **(**$p < 0.05$**)**

5.2.5 多重共线性检查

为了确保后续回归模型的稳健性与解释力，我们对所有数值型预测变量进行了多重共线性检查，并计算了方差膨胀因子（VIF）。VIF 值能够直观反映一个自变量能否被其他自变量线性解释的程度，从而判断是否存在过度相关性。如图 9 所示，教育程度 _ 研究生的 VIF 值达到 5.72，这是所有变量中最高的。尽管该值接近但尚未超过通常设定为 10 的警戒线，这表明教育程度的某些类别之间存在一定程度的共线性，但其影响仍在模型可接受的范围内。类似地，教育程度 _ 高中和教育程度 _ 大学的 VIF 值也相对较

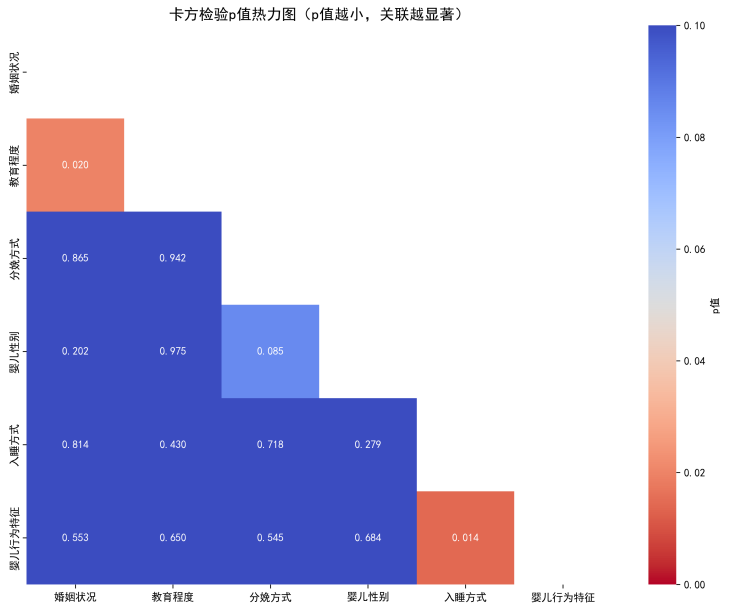

图 **8**　卡方检验 **p** 值热力图 **(p** 值越小，关联越显著**)**

高，分别为 4.56 和 4.36，进一步印证了教育背景变量内部的关联性。母亲 EPDS 得分的 VIF 值为 3.94，这反映了 EPDS 与其他心理指标之间存在一定的相关性，与我们之前在相关性分析中的发现相符。值得庆幸的是，绝大多数预测变量的 VIF 值均小于 2.5，这有力地表明模型中多重共线性问题得到了有效控制，预计不会对模型结果的稳定性或参数估计的准确性产生严重影响。

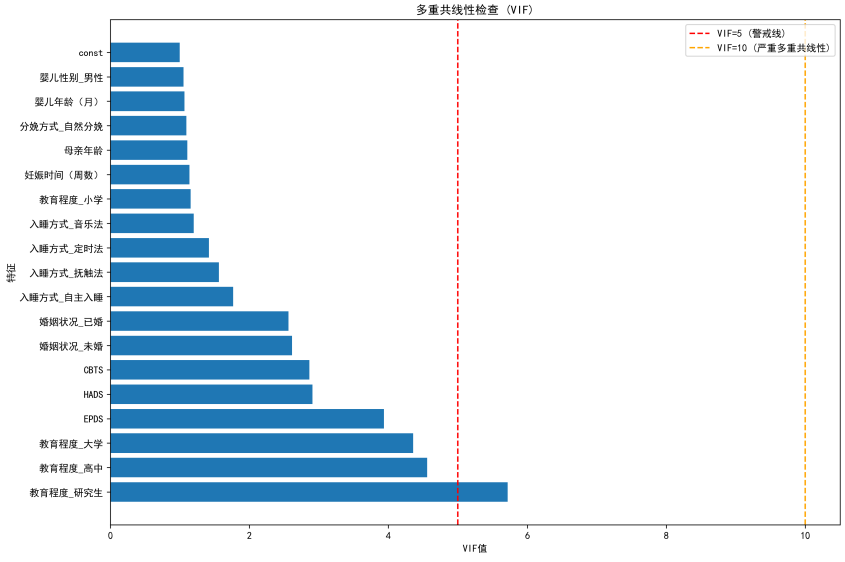

图 **9**　多重共线性检查 **(VIF)**

5.3 模型选择与理论基础

根据问题一所涉因变量的类型（连续型与多分类型），我们选择了相应的统计回归模型进行分析，以期最大化地捕捉变量间的潜在关系。

5.3.1 婴儿睡眠质量模型：多元线性回归

鉴于婴儿的“整晚睡眠时间”和“睡醒次数”均为连续数值型变量，我们选择多元线性回归 (Multiple Linear Regression) 作为其预测模型。多元线性回归旨在建立一个线性关系，用一个或多个自变量来预测因变量的取值，其优点在于模型简洁且易于解释。其一般形式为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

其中：Y 是因变量（如整晚睡眠时间或睡醒次数）。$X_1, X_2, \ldots, X_k$ 是 k 个自变量（如母亲年龄、心理指标、入睡方式等）。β_0 是截距项。$\beta_1, \beta_2, \ldots, \beta_k$ 是对应自变量的回归系数，表示在其他自变量不变的情况下，该自变量每增加一个单位，因变量的平均变化量。ϵ 是误差项，代表模型未能解释的随机变异。模型通过最小化残差平方和（Ordinary Least Squares, OLS）来估计回归系数。模型的评估指标包括决定系数 (R^2)、调整决定系数 (Adjusted R^2)、F 统计量及其 p 值，用于评估模型的整体拟合优度和统计显著性。

5.3.2 婴儿行为特征模型：多项逻辑回归

婴儿行为特征（“安静型”、“中等型”、“矛盾型”）是一个具有三个或更多无序类别的分类变量。因此，我们选择多项逻辑回归 (Multinomial Logistic Regression) 模型来分析母亲各项因素对其影响。多项逻辑回归是二元逻辑回归的自然扩展，用于预测一个多分类因变量的概率，避免了将无序类别强行排序可能引入的偏差。它通过建立多个二元逻辑回归模型来比较每个类别与一个基准类别（参考类别）的对数发生比 (log-odds)。假设我们有 J 个类别，选择第 J 个类别作为参考类别。对于每个非参考类别 $j \in \{1, \ldots, J-1\}$，模型估计其相对于参考类别的对数发生比：

$$\ln\left(\frac{P(Y=j|\mathbf{X})}{P(Y=J|\mathbf{X})}\right) = \beta_{j0} + \beta_{j1} X_1 + \beta_{j2} X_2 + \cdots + \beta_{jk} X_k$$

其中：$P(Y=j|\mathbf{X})$ 是在给定自变量 $\mathbf{X}$ 的情况下，因变量属于类别 j 的概率。β_j 是对应于类别 j 的回归系数向量。发生比 (Odds Ratio, OR)：$e^{\beta_{ji}}$ 表示在其他自变量不变的情况下，自变量 X_i 每增加一个单位，因变量属于类别 j 而而非参考类别的发生比的变化倍数。模型的评估通常使用赤池信息准则 (AIC) 和贝叶斯信息准则 (BIC)，它们平衡了模型的拟合优度和复杂度。发生比 (Odds Ratio) 的可视化是理解自变量对不同类别影响的关键，能够直观展示各因素对不同行为类型的倾向性影响。

5.4 模型构建与求解

在确定模型类型后，我们基于前期探索性数据分析的结果，构建并求解了各个模型，以量化母亲因素对婴儿睡眠与行为的影响。

5.4.1 婴儿睡眠质量线性回归模型构建

我们分别构建了预测“整晚睡眠时间”和“睡醒次数”的线性回归模型。模型中纳入了所有经过预处理的母亲身体指标、心理指标以及独热编码后的分类变量。

整晚睡眠时间模型 该睡眠时间模型在统计学上整体显著（$F = 4.66, p < 0.001$），这表明模型中的自变量组合能够对婴儿的整晚睡眠时间产生显著影响。然而，模型的解释力相对有限，其决定系数 R^2 为 0.184，调整决定系数 $Adjusted\ R^2$ 为 0.145。这提示虽然模型捕捉到了一部分变异，但婴儿睡眠时间的影响因素可能更为复杂多样，现有变量仅能解释约 18.4% 的变异，可能存在更多未被纳入的生理或环境因素。在显著预测因子方面，入睡方式中的“自主入睡”表现出最为显著的正向影响。这意味着，与其他入睡方式相比，能够自主入睡的婴儿普遍拥有显著更长的整晚睡眠时间。这一发现不仅再次验证了前期探索性数据分析（ANOVA）的结论，更强调了培养婴儿自主入睡能力对提升其睡眠质量具有决定性的作用。此外，母亲的 EPDS（爱丁堡产后抑郁量表）得分与婴儿整晚睡眠时间呈现显著的负相关关系。具体而言，母亲的 EPDS 得分越高，其婴儿的整晚睡眠时间则越短。这一结果清晰地揭示了母亲抑郁情绪对婴儿睡眠模式的负面影响，从而提示了心理干预的必要性。

睡醒次数模型 与睡眠时间模型类似，睡醒次数模型在统计上整体显著（$F = 5.32, p < 0.001$），且其解释力略高于睡眠时间模型（$R^2 = 0.205$, Adjusted $R^2 = 0.167$）。在显著预测因子方面，如同在睡眠时间模型中，入睡方式在此模型中依然是影响婴儿夜间睡醒次数最为显著的因素。具体分析显示，自主入睡的婴儿夜间睡醒次数明显更少，进一步强调了其对改善婴儿睡眠连续性的积极作用，并为育儿指导提供了明确方向。此外，母亲的 HADS（医院焦虑抑郁量表）得分与婴儿夜间睡醒次数呈现显著的正相关关系。这意味着母亲的焦虑抑郁程度越高，婴儿夜间睡醒的次数越多。这一结果有力地印证了母亲心理健康状况与婴儿睡眠质量之间的紧密联系，提示了心理压力对婴儿睡眠节律的干扰。

5.4.2 婴儿行为特征多项逻辑回归模型构建

考虑到婴儿行为特征（“安静型”、“中等型”、“矛盾型”）属于无序多分类变量，我们构建了多项逻辑回归模型以探究母亲各项因素对其影响。在此模型中，我们选择将“中等型”婴儿作为参考类别，进而分析其他两类（“安静型”和“矛盾型”）相对于“中

等型”的对数发生比。模型的评估指标分别为 $AIC = 693.31$ 和 $BIC = 844.02$。尽管模型能够有效区分不同的行为特征类型，但其预测准确率仍存在进一步优化的空间。在显著预测因子与发生比分析方面，模型结果显示，CBTS、EPDS 和 HADS 这三项母亲心理健康指标对婴儿行为特征具有显著的影响力。具体而言，当母亲的心理压力越大，即 CBTS、EPDS 和 HADS 得分越高时，其婴儿表现为“矛盾型”行为特征的可能性相对于“中等型”婴儿显著增加。这一发现强烈提示，母亲的心理健康问题可能直接导致或加剧婴儿出现更具挑战性的行为模式。入睡方式被证实是影响婴儿行为特征的另一个关键因素。相较于“中等型”婴儿，采用“自主入睡”方式的婴儿更倾向于表现为“安静型”行为特征。这进一步强调了培养良好入睡习惯的重要性，更揭示了其在塑造婴儿积极行为模式方面的潜在作用。母亲的教育程度对婴儿行为特征也展现出一定的影响。通常观察到，教育程度越高的母亲，其婴儿表现出“安静型”行为特征的可能性越大。这可能间接反映了较高教育程度的母亲在育儿理念、家庭环境创设或早期干预策略上的优势，从而有利于培养婴儿的稳定行为模式。图 10 直观展示了各预测变量对婴儿行为特征影响的发生比，为上述分析提供了可视化证据。

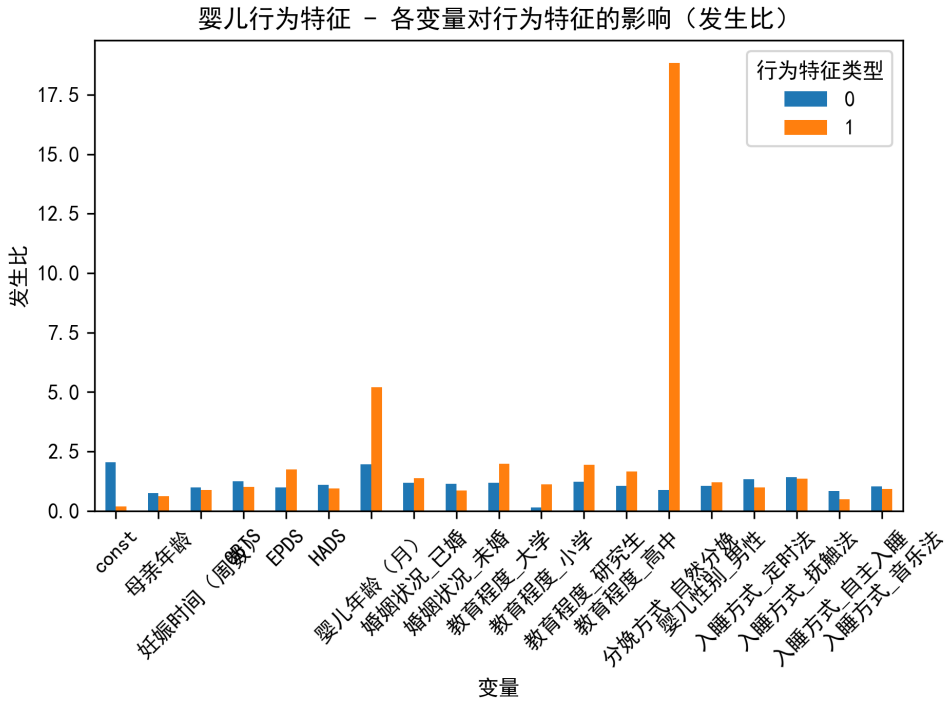

图 **10**　婴儿行为特征发生比可视化

5.5 模型结果与分析

综合上述多元线性回归与多项逻辑回归模型的分析结果，我们得以深入剖析母亲身心健康与婴儿早期发展之间的复杂关系，并得出以下关键发现：

本研究构建的模型一致揭示，母亲的抑郁情绪（EPDS）、焦虑水平（HADS）以及分娩相关创伤后应激（CBTS）症状与婴儿的睡眠质量及行为特征之间存在显著且不容忽视的关联。具体而言，母亲所承受的心理压力越大，其婴儿的整晚睡眠时间越短，夜间睡醒次数越多，并且其行为模式也越倾向于表现出“矛盾型”特征。这一核心发现有力地强调了在产妇保健体系中，将心理健康筛查与早期干预纳入常规服务的重要性，以积极促进婴儿的全面健康发展。

无论是婴儿的整晚睡眠时间、夜间睡醒次数的分析中，还是在婴儿行为特征的预测模型里，入睡方式均被反复确认为最关键的影响因素之一。研究结果清晰表明，那些能够自主入睡的婴儿普遍拥有更长的整晚睡眠时间、更少的夜间睡醒次数，并且其行为模式更倾向于表现为“安静型”特征。这不仅为日常育儿实践提供了明确且具有操作性的指导，更凸显了积极引导和培养婴儿自主入睡能力在改善其睡眠质量和促进其良好行为模式发展方面的核心价值。

尽管本研究构建的模型在统计学上均达到了显著水平，但多元线性回归模型所能解释的婴儿睡眠质量变异比例（R^2 值约为 0.18 至 0.20）相对有限。这提示我们，婴儿的睡眠和行为发展是一个极其复杂的生物-心理-社会交互过程，除了本模型所纳入的母亲身心健康状态和育儿方式等可量化因素外，很可能还存在大量未被模型捕捉或量化的遗传因素、环境影响、生理机制或其他社会经济因素，这些未纳入的因素共同作用，导致了婴儿发展中显著的个体差异。因此，在实际的临床干预和育儿指导中，必须充分考虑个体的特异性，并提供多维度、个性化的支持方案。

母亲的 CBTS、EPDS 和 HADS 得分之间表现出高度的相关性，这强烈暗示了这些心理困扰并非孤立存在，而是倾向于相互关联并同时出现的综合性问题。在多项逻辑回归模型中，这些心理指标共同作用于婴儿的行为特征，进一步凸显了它们对婴儿发展模式的综合性影响。这一发现进一步强调了在临床实践中，对产妇心理健康的干预应采取更为综合、整体性的策略，而非仅仅针对单一症状进行碎片化处理，以期达到更全面的积极效果。

六、 问题二的模型的建立和求解

6.1 模型建立

本题旨在建立婴儿行为特征（安静型、中等型、矛盾型）与母亲身体及心理指标的关系模型，并预测未知样本类别。基于前述分析，模型流程如下：

6.1.1 模型选择与模型结构：

随机森林（Random Forest）是本题的最佳选择。理由如下：

1. **高精度与鲁棒性：**作为一种集成学习方法，它通过构建大量独立决策树并进行多数投票，显著提高了预测精度和抗噪声能力，不易过拟合。
2. **处理非线性关系：**决策树本身能够捕捉特征与目标之间复杂的非线性关系和潜在的交互作用，这在复杂的医学/心理学数据中尤为重要。
3. **特征重要性：**随机森林能够直接输出每个特征的重要性，这为我们提供了模型可解释性，有助于识别影响婴儿行为特征的关键母亲指标。
4. **对类别不平衡的缓解：**虽然不是根本解决，但结合 `class_weight='balanced'` 参数，随机森林能在一定程度上缓解类别不平衡问题。

模型结构：

对于每一棵决策树 $h_b(\cdot)$，$b = 1, \ldots, B$，其在样本 $\mathbf{x}$ 上的输出为 $\hat{y}_b = h_b(\mathbf{x})$。集合所有树的输出，分类结果为

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_B(\mathbf{x})\} \tag{1}$$

其中 mode$\{\cdot\}$ 为众数操作。

6.1.2 数据预处理

目标：将原始数据转换为适用于机器学习模型的结构和格式，并进行训练集与预测集的划分。

变量识别：

- **自变量 (特征)** X：母亲年龄、婚姻状况、教育程度、妊娠时间（周数）、分娩方式、CBTS、EPDS、HADS。
- **因变量 (目标)** Y：婴儿行为特征（安静型、中等型、矛盾型）。
- **辅助变量：**编号 (用于关联预测结果)。

步骤：

1. **数据加载与分离：**
   - 从 数据`.csv` 文件加载原始数据。
   - 根据 婴儿行为特征列是否为空或空字符串，将原始数据划分为：
     - **训练集 (Training Set):** 包含 婴儿行为特征值的 390 条记录，用于模型训练和评估。
     - **待预测集 (Prediction Set):** 婴儿行为特征为空的 20 条记录（编号 391-410），用于最终预测。
2. **目标变量编码：**
   将训练集中的 婴儿行为特征列进行数值编码，转换为模型可识别的整数标签。其中：安静型 $\to 0$, 中等型 $\to 1$, 矛盾型 $\to 2$,

$$Y_{\text{encoded},i} = \begin{cases} 0 & \text{if 婴儿行为特征}_i = \text{安静型} \\ 1 & \text{if 婴儿行为特征}_i = \text{中等型} \\ 2 & \text{if 婴儿行为特征}_i = \text{矛盾型} \end{cases}$$

3. **类别分布分析：**
   - 统计并报告训练集中三种婴儿行为特征的**具体样本数量和百分比**。
   - **强调不平衡性：**明确指出“矛盾型”为少数类别，这将在后续模型构建和评估中重点关注。
4. **特征类型分类：**
   - 明确区分输入特征中的**数值型特征**（母亲年龄、妊娠时间（周数）、CBTS、EPDS、HADS）。
   - 明确区分**类别型特征**（婚姻状况、教育程度、分娩方式）。
5. **数据转换管道构建：**
   - 使用 `sklearn.compose.ColumnTransformer` 构建预处理管道。
     - 对于数值型特征，应用 **StandardScaler** 进行标准化。
     - 对于类别型特征，应用 **OneHotEncoder** 进行独热编码。
     - **关键参数：**设置 `handle_unknown='ignore'`，以防止在预测阶段遇到训练集中未出现的新类别时引发错误。
   - **数学表达式（标准化）：**对于任意数值特征 X_j，其标准化后的值 $X'_j = \frac{X_j - \bar{X}_j}{s_j}$，其中 $\bar{X}_j$ 和 s_j 分别为训练集中 X_j 的均值和标准差。
   - **理由：**这种集成预处理的方法确保了后续建模过程的自动化、一致性，并有效避免了数据泄露。
6. **训练集与验证集划分：**
   - 将已编码的训练集（390 条记录）进一步划分为**训练子集**（例如 80%）和**验证集**（20%）。
   - **关键方法：**采用**分层抽样**（`stratify=y_train_encoded`），确保训练子集和验证集中各行为特征类别的比例与原始训练集保持一致。这对于评估模型在不同类别上的真实性能至关重要。

6.1.3 模型管道定义：

- 构建一个 `sklearn.pipeline.Pipeline` 对象,将数据预处理步骤（`ColumnTransformer`）和随机森林分类器（`RandomForestClassifier`）串联起来。
- **示例：**

```
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor_object),
    ('classifier', RandomForestClassifier(random_state=
   Config.RANDOM_STATE))
])

```

6.1.4 超参数调优

目标：通过系统搜索，找到使随机森林模型性能最优的超参数组合，以提高其在未知数据上的泛化能力。

步骤如下：

1. **调优方法：**采用**网格搜索（GridSearchCV）**进行超参数优化。
2. **参数网格定义：**明确定义随机森林分类器的超参数搜索空间。
   - `classifier__n_estimators`: 决策树的数量，例如 `[100, 200, 300]`。树越多模型越稳定，但计算成本越高。
   - `classifier__max_features`: 每次分裂时考虑的最大特征数，例如 `['sqrt', 'log2', 0.8]`。控制决策树的多样性。
   - `classifier__max_depth`: 决策树的最大深度，例如 `[None, 10, 20]`。限制树的生长，防止过拟合。
   - `classifier__min_samples_split`: 分裂内部节点所需的最小样本数,例如 `[2, 5]`。
   - `classifier__min_samples_leaf`: 叶子节点所需的最小样本数，例如 `[1, 2]`。
   - `classifier__class_weight`: 类别权重,例如 `[None, 'balanced']`。`'balanced'` 参数可自动根据类别样本比例调整权重，以减轻类别不平衡对少数类预测的影响。
3. **交叉验证策略：**在 `GridSearchCV` 中使用 **K 折交叉验证**（例如 $K = 5$）。
   - **理由：**交叉验证能够更全面地评估模型性能，减少模型评估结果对特定训练/测试集划分的依赖，提高模型评估的可靠性。
4. **评估指标：**指定 `scoring='f1_weighted'` 作为 `GridSearchCV` 的优化目标。
   - **理由：**加权 F1 分数对类别不平衡数据更敏感，能更客观地反映模型在所有类别上的综合性能，避免模型仅仅在多数类别上表现良好。
5. **计算资源：**配置 `n_jobs=-1` 以利用所有可用 CPU 核心进行并行计算，加速调优进

程。

6. **结果报告：**报告 `GridSearchCV` 找到的**最佳超参数组合**（`.best_params_`）和对应的**最高交叉验证加权 F1 分数**（`.best_score_`）。

模型框架如 (2) 所示。

$$f^* = \arg\max_{f \in \mathcal{F}} \; F_1^{\text{weighted}} \tag{2}$$

本方法参考了文献 [**?**] 关于分类模型的设计思想，并借鉴了[?] 中基于管道的机器学习实现。

6.2 模型求解

Step 1：数据划分与特征处理

首先，将原始数据根据“婴儿行为特征”是否缺失分为训练集（390 条有标签）与预测集（20 条无标签）。对母亲年龄、妊娠时间、CBTS、EPDS、HADS 等数值型变量进行标准化，对婚姻状况、教育程度、分娩方式等类别型变量特征进行标准化和独热编码处理，目标变量完成 0-1-2 数值化编码。所有预处理操作在 Pipeline 中自动完成。

对训练集类别进行分析。其分布见图11：可以看出样本分布中，中等型数量最多，

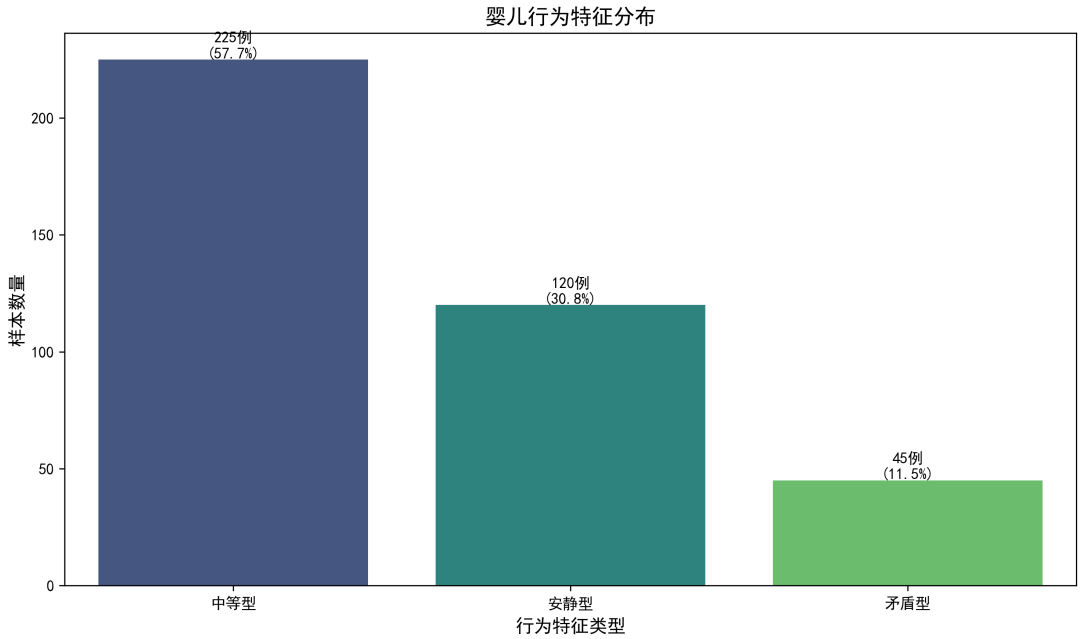

图 **11**　婴儿行为特征分布（训练集）

矛盾型最少，数据呈现明显类别不平衡。

Step 2：模型训练与超参数调优

采用带有预处理的随机森林分类器建模。通过 GridSearchCV，使用训练集进行随机森林建模，通过网格搜索（GridSearchCV）自动寻优参数组合，设置 5 折交叉验证并采用加权 F1 作为指标，缓解类别不平衡影响。在如下参数空间内寻优：

- 决策树数量 $n_{\text{estimators}} \in \{100, 200, 300\}$；
- 最大树深 $max\_depth \in \{\text{None}, 10, 20\}$；
- 分裂最小样本 $min\_samples\_split$ 等；
- 类别权重 $[None, 'balanced']$，选择能缓解类别不均的权重参数；
- 各参数组合以加权 F1 分数为优化目标。

最终获得最优模型，将其在训练集上训练，并利用验证集评估性能。

Step 3：模型评价与可视化

模型在测试集上的主要评价指标如下：

$$\text{准确率} = 0.5513 \qquad \text{加权 } \mathrm{F}_1 = 0.4664$$

具体精度、召回率和 F_1 指标如表，说明模型对“中等型”婴儿识别能力较好，对“矛盾型”能力较弱。

进一步，利用混淆矩阵分析模型对各类别的判别能力（见图12）：

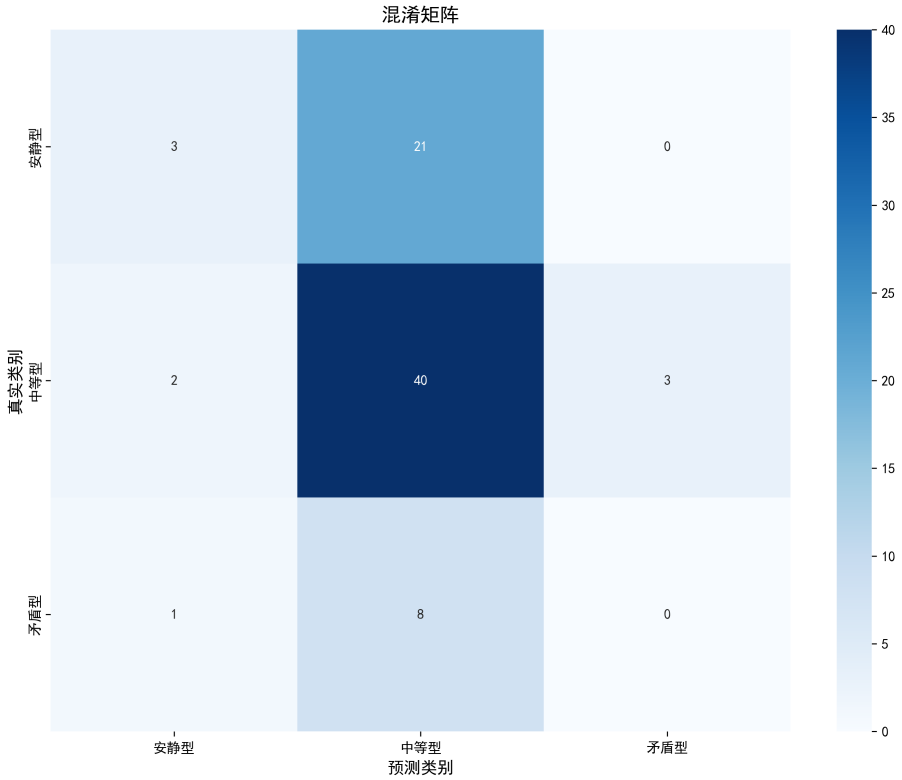

图 **12**　模型预测的混淆矩阵

由图可见，“中等型”易被正确分类，而“矛盾型”大都被误判为“中等型”，反映出少数类别的预测困难。

Step 4：特征重要性分析

训练完成后，输出各特征对模型的重要性评分，并作排序可视化（见图13）：可以

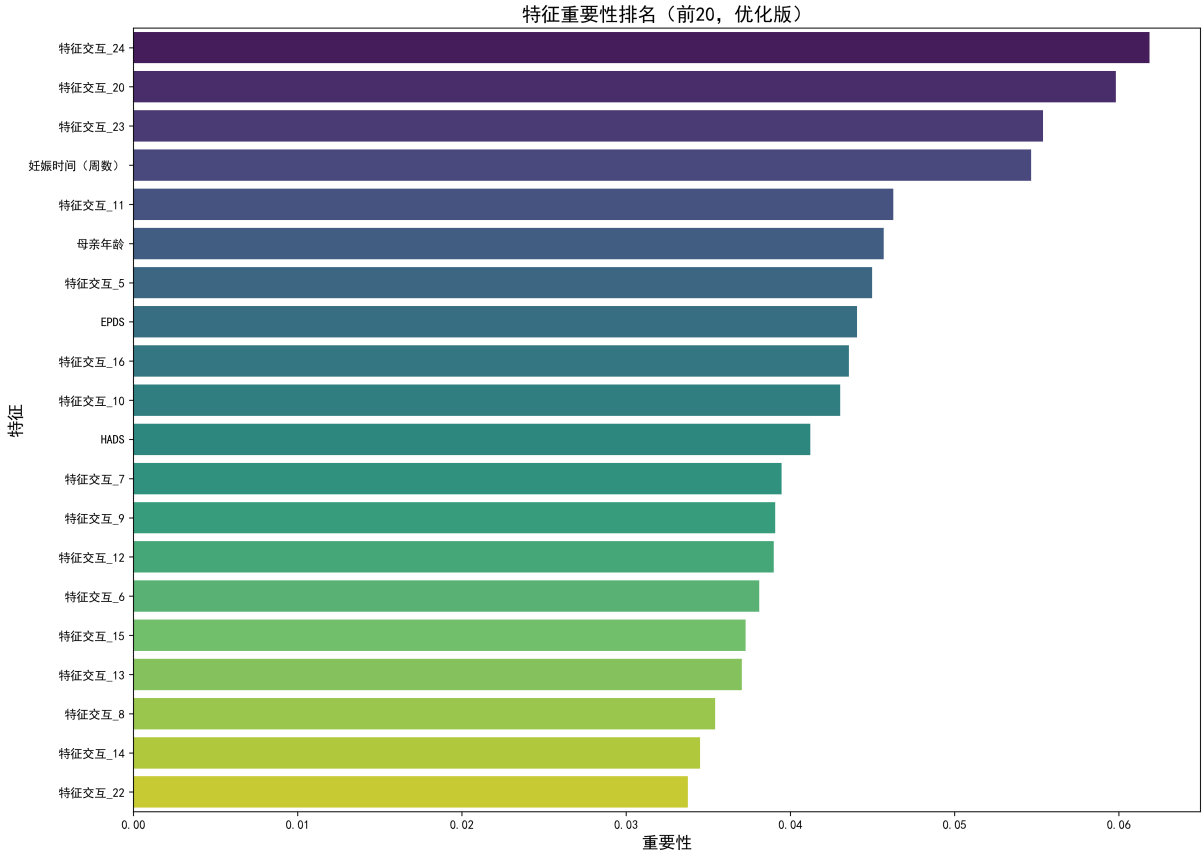

图 **13**　特征重要性排名（部分）

看出，妊娠时间（周数）、母亲年龄及心理健康指标（EPDS、HADS、CBTS 等）在婴儿行为预测中起到主导作用。这些关键特征的识别为后续的干预策略建议提供专业依据。

Step 5：未知样本预测及分布

最终，利用全体训练数据和最优模型，预测编号 391-410 的 20 个婴儿行为特征。预测结果分布见图14：

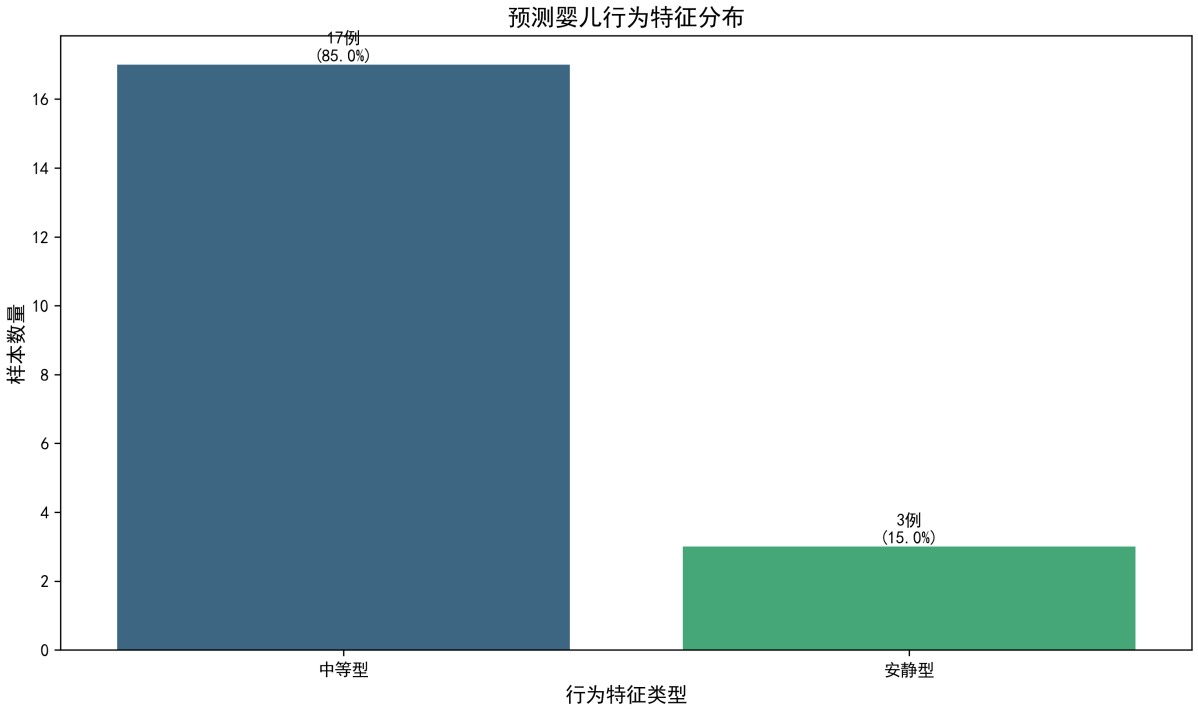

图 **14**　预测婴儿行为特征分布（编号 **391-410**）

可见“中等型”预测概率最高，矛盾型未被预测。这与模型在验证集上的表现一致，进一步说明类别不平衡对模型泛化能力的影响。

Step 6：小结与讨论

模型有效实现了婴儿行为特征分类与未知样本预测，能够定量输出特征重要性。预测表现受限于类别不平衡和样本量有限，未来可考虑数据扩充、过采样及提升模型结构等改进方案。

6.3 求解结果

本节具体阐述模型训练、评估与预测全过程，包含数据分析、建模、评估、特征解释及结果预测。

- **整体性能：**测试集准确率 0.5513，加权 F1 为 0.4664。
- **分类表现：**模型对中等型婴儿识别最优（F1=0.70），对安静型次之（F1=0.20），对矛盾型识别能力较弱（F1=0）。
- **特征重要性：**妊娠时间、EPDS、母亲年龄、HADS、CBTS 等为影响婴儿行为特征的关键因子。
- **预测分布：** 20 例未知婴儿预测结果为中等型 17 例，安静型 3 例，矛盾型 0 例，推论模型对少数类仍存失衡问题。

分析结果表明，提升对少数类（如矛盾型）识别能力仍是今后工作的重点，后续可尝试采用 SMOTE 等过采样、XGBoost 等进一步优化模型泛化与公平性。

七、 问题三的模型的建立和求解

7.1 模型建立

7.2 模型求解

Step1：

Step2：

Step3：

7.3 求解结果

八、 问题四的模型的建立和求解

8.1 模型建立

8.2 模型求解

Step1：

Step2：

Step3：

8.3 求解结果

九、 模型的分析与检验

9.1 灵敏度分析

9.2 误差分析

十、 模型的评价

10.1 模型的优点

- 优点 1
- 优点 2
- 优点 3

10.2 模型的缺点

- 缺点 1
- 缺点 2

附录 A　文件列表

文件名	功能描述
q1.m	问题一程序代码
q2.py	问题二程序代码
q3.c	问题三程序代码
q4.cpp	问题四程序代码

附录 B　代码

q1.m

```m
disp("Hello World!")
```

q2.py

```python
print("Hello World!")
```

q3.c

```c
#include <stdio.h>

int main()
{
    printf("Hello World!");
    return 0;
}
```

q4.cpp

```cpp
#include <bits/stdc++.h>
using namespace std;

int main()
{
    cout << "Hello World!" << endl;
    return 0;
}
```