基于XXX模型的XXX问题研究

摘要

（1）标题一般写成：**基于XXX模型/方法/理论的XXX问题研究**

（2）摘要一般包括三部分内容：**前言、正文和结尾**

（3）摘要是整个论文**最重要的部分**，一定认真准备，不要超过一页

（4）摘要内容概括：摘要前言主要起到总结概括的作用，一般交代一下背景和主要解决的问题即可，一般写3-5行。摘要正文主要写针对每一小问的建模过程和求解思路。摘要收尾是对整个建模过程的总结和升华，常见的是进行优缺点评价、模型的创新性评价、模型的推广等

（5）关键词：一般写研究对象、主要模型或求解算法等，3-5个即可。

总体内容形式：

开头段：本文针对XXX问题，主要建立了XXX模型，求解得到XXX结果。

针对问题一，主要解决XXX问题，通过XXX分析，建立XXX模型，基于XXX方法/软件等，求解得到XXX结果。

针对问题二，主要解决XXX问题，通过XXX分析，建立XXX模型，基于XXX方法/软件等，求解得到XXX结果。

针对问题三，主要解决XXX问题，通过XXX分析，建立XXX模型，基于XXX方法/软件等，求解得到XXX结果。

摘要收尾：是对整个建模过程的总结和升华，常见的是进行优缺点评价、模型的创新性评价、模型的推广等。

关键词：XXX，XXX，XXX，XXX

**目录**

[基于XXX模型的XXX问题研究 1](#_Toc18754)

[摘要 1](#_Toc22414)

[一、 问题重述 3](#_Toc27344)

[二、 问题分析 3](#_Toc28431)

[2.1 问题一的分析 3](#_Toc32494)

[2.2 问题二的分析 4](#_Toc13671)

[2.3 问题三的分析 4](#_Toc27630)

[三、 模型假设 4](#_Toc22007)

[四、 符号说明 4](#_Toc23726)

[五、 模型建立与求解 5](#_Toc28934)

[5.1 问题一模型的建立与求解 6](#_Toc5559)

[5.1.1 数据预处理 6](#_Toc12522)

[5.1.2 XXX模型的建立 6](#_Toc30769)

[5.1.3 XXX模型的求解 6](#_Toc31148)

[5.1.4 XXX结果的分析 6](#_Toc18808)

[5.2 问题二模型的建立与求解 6](#_Toc3788)

[5.3 问题三模型的建立与求解 6](#_Toc19545)

[六、 模型检验 6](#_Toc10243)

[七、 模型优缺点评价 6](#_Toc24900)

[7.1 模型的优点 6](#_Toc30007)

[7.2 模型的缺点 7](#_Toc19695)

[7.3 模型的改进 7](#_Toc21249)

[参考文献 8](#_Toc326)

[附录 9](#_Toc15480)

# 问题重述

1.1问题背景

随着产前筛查技术的发展以及精准医疗概念的普及，无创产前检测（NIPT）已成为早期识别胎儿健康隐患的重要手段。该技术通过采集母体血液、提取胎儿游离 DNA 片段从而分析其染色体是否存在异常。根据临床实践可知，胎儿染色体异常主要集中于三类综合征：唐氏综合征（由 21 号染色体游离 DNA 片段比例异常引发）、爱德华氏综合征（18 号染色体浓度异常）与帕陶氏综合征（13 号染色体浓度异常），而 NIPT 检测结果的准确性，核心取决于胎儿性染色体浓度：若男胎 Y 染色体浓度达到或超过 4%、女胎 X 染色体浓度无异常，检测结果可满足临床基本准确性要求。

临床已明确分级标准：12 周以内（早期）发现异常时，医疗干预风险较低；13-27 周（中期）发现时，干预难度与风险显著上升；28 周以后（晚期）发现时，已错过最佳干预阶段，风险极高。而 NIPT 检测窗口期恰好覆盖 “早期——中期” 的关键阶段，因此 “时点选择” 需严格匹配风险分级，避免因时点不当缩短治疗窗口期。已有临床研究表明，男胎的 Y 染色体浓度不仅与孕周有关，还受到孕妇体重指数（BMI）的显著影响。由于不同孕妇的个体差异，采用统一的检测时点可能导致部分孕妇的检测不准确。因此，研究如何根据孕妇的 BMI 及其他特征确定最佳检测时点，对提高检测准确性、降低潜在风险具有重要意义。

1.2问题提出

基于附件给出的某地区孕妇（大多BMI较高）的NIPT数据，建立数学模型从而解决以下问题：

问题一：针对男胎孕妇群体，剖析胎儿 Y 染色体浓度与孕妇孕周、BMI 等关键指标间的关联特征，构建能定量描述这些关系的数学模型，并通过统计方法验证模型的显著性，确保模型对变量关系刻画的可靠性。

问题二：已知孕妇 BMI 是影响男胎 Y 染色体浓度达标时间的核心因素，需基于此对男胎孕妇进行科学分组，明确每组的 BMI 区间范围，并确定各组的最佳 NIPT 检测时点。该时点应当以 “最小化孕妇潜在风险” 为目标，同时分析检测误差对分组结果及时点选择的影响，评估模型稳健性。

问题三：男胎 Y 染色体浓度达标时间受多重因素共同作用，包括孕妇身高、体重、年龄等，且需考虑检测误差与 Y 染色体浓度达标比例。需综合这些因素，再次对男胎孕妇按 BMI 进行合理分组，确定各组最佳 NIPT 时点以最小化潜在风险，并进一步分析检测误差对最终结果的干扰程度，完善时点选择方案。

问题四：由于孕妇与女胎均不携带 Y 染色体，女胎异常判定需关注染色体非整倍体。需整合 X 染色体及 21 号、18 号、13 号染色体的 Z 值、GC 含量、读段数及相关比例、孕妇 BMI 等多维度数据，提出女胎异常判定方法。

# 问题分析

## 问题一的分析

问题一要求利用已知数据集，分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的定量关联，构建关系模型并检验显著性，为后续 BMI 分组及最佳 NIPT 时点选择提供理论依据。考虑到分析需排除无关样本干扰且数据集中包含非线性变量，先利用数据预处理方法筛选有效样本与规范变量，即仅保留男胎样本、将孕周字符串转换为连续数值、将日期转换为标准日期格式。

数据预处理后，通过散点图观察 Y 染色体浓度与孕周、BMI 的关系。计算 Pearson 相关系数初步衡量线性关联强度。若变量间线性关联显著，则使用多元线性回归模型量化影响幅度；反之则数据预处理后，通过散点图观察 Y 染色体浓度与孕周、BMI 的关系。计算 Pearson 相关系数初步衡量线性关联强度。若变量间线性关联显著，则使用多元线性回归模型量化影响幅度；反之则使用广义加性混合模型 （GAMM），从而刻画变量之间的非线性关联。模型拟合完成后，利用统计检验与显著性分析，确定孕周与 BMI 对 Y 染色体浓度的主要影响因素及其贡献，为后续的 BMI 分组与最佳检测时点选择提供依据。

## 问题二的分析

问题二旨在探讨孕妇 BMI 对男胎 Y 染色体浓度达标孕周的影响，并据此确定最佳 NIPT 检测时点。由于 BMI 是影响 Y 染色体浓度达标时间的主要因素，首先对样本进行 BMI 分组，采用条件推理树将孕妇划分为若干区间，以保证组内样本均衡。

随后，在各 BMI 组内利用 Kaplan–Meier 方法估计累计达标率曲线，寻找使达标率达到 95% 的孕周作为推荐检测时点；若组内最高达标率不足 95%，则以其最大值对应的孕周作为替代。为验证结果的稳健性，进一步采用 Bootstrap 方法进行区间估计，并引入 Logistic 回归对孕周与达标概率关系进行平滑拟合，与 KM 曲线进行对比。最终通过表格与曲线输出各 BMI 组的推荐检测时点及置信区间，结果表明 BMI 越高，推荐检测孕周越晚。该分析揭示了 BMI 对 NIPT 检测时点选择的分层效应，为临床合理安排检测提供了科学依据。

## 问题三的分析

## 问题四的分析

# 模型假设

对情景的说明，不可能也不必要提供问题的每个细节。由此而来建立由此而来建立数学模型还是不够的，还要补充一些假设，模型假设是建立数学模型中非常关键的一步，关系到模型的成败和优劣。

常见的假设如下：

**① 对题目中已知条件或参数做出保真性假设**

如：题目给出的测量数据准确，无异常值；

**② 仅考虑题目中涉及的主要条件，对其他情况不考虑或进行强制规定**

如：不考虑高压油管内壁对油的粘滞力;

**③ 对题目中涉及的主要条件进行平稳性规定**

如：整个系统温度恒定

**④ 为使研究更简便、或从常识性角度做出的假设**

如：以第四层(空气层)底层温度表示人体皮肤外侧温度

**⑤ 对模型中相关参数做出规定**

如：人体为绝对黑体，即辐射发射率为1

# 符号说明

表1 本文的符号说明

|  |  |  |
| --- | --- | --- |
| **符号** | **说明** | **单位** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  | / |
|  |  |  |
|  |  |  |
|  |  |  |

# 模型建立与求解

## 问题一模型的建立与求解

### 数据预处理

**在建立模型、求解问题前，首先应当对数据进行预处理。读取原始数据文件并对主要字段进行初步检查。针对“孕妇BMI”、“Y染色体浓度”、“检测孕周”等关键变量，统计并筛查缺失值和异常值。其次，对时间相关变量进行标准化处理。将“末次月经”和“检测日期”字段统一转换为标准日期格式，便于后续孕周计算及时间相关分析。接着，对“检测孕周”变量进行量化处理，采用正则表达式提取孕周和天数，并统一转换为以周为单位的数值型变量。例如，“12w+3”转换为 12 + 3/7 ≈ 12.43 周，并保留两位小数。对于缺失或无效格式的数据，统一填充为缺失值。**

### Pearson相关性分析

为探究孕妇BMI、检测孕周与Y染色体浓度之间的相关性，采用Pearson相关系数进行分析。计算变量x与y的Pearson相关系数计算公式如下：

|  |  |
| --- | --- |
|  | （1） |

选取孕妇BMI、检测孕周、Y染色体浓度三者作为分析对象，利用pandas和scipy库计算三者之间的Pearson相关系数，可以绘制出相关性热力图如下图所示：

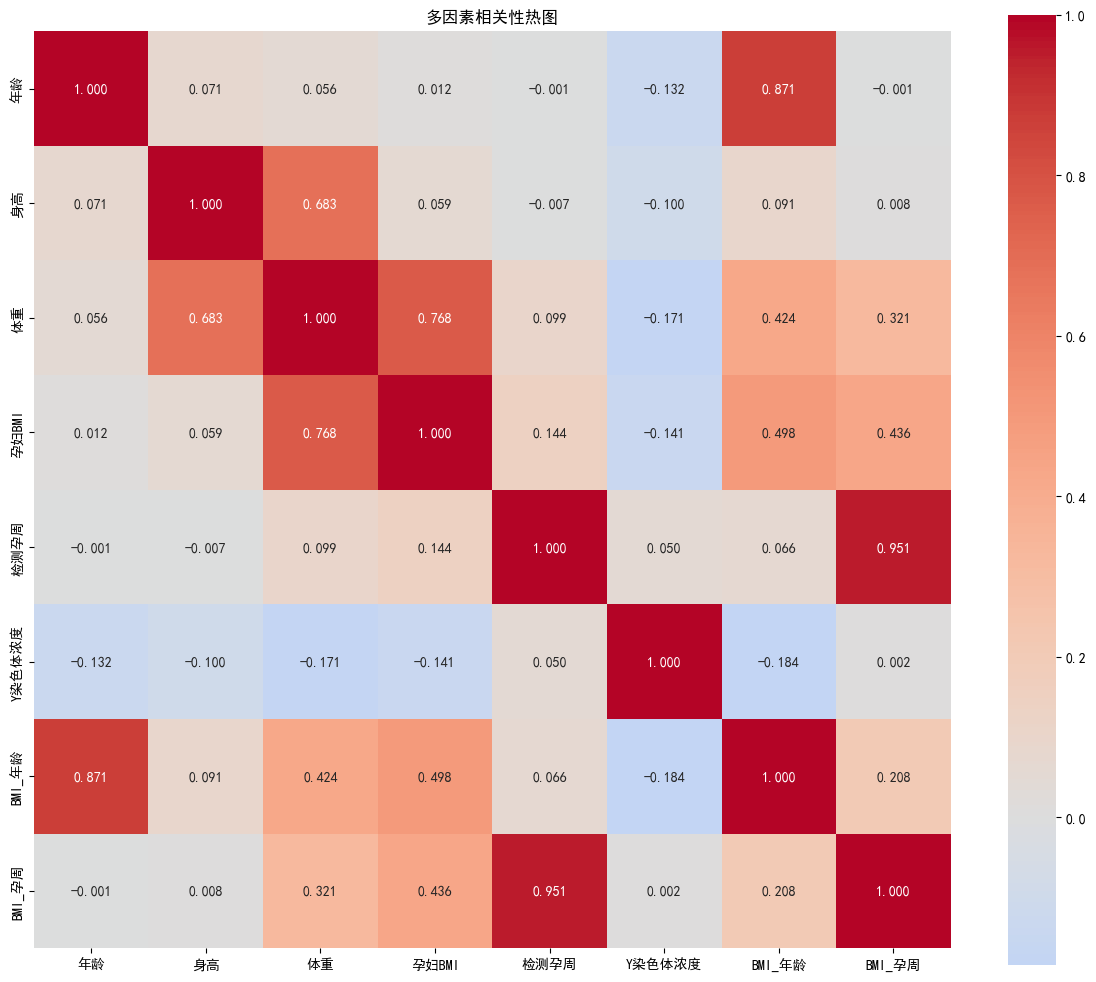


图1 相关系数矩阵热力图

**Pearson相关系数值域为[-1,1]，其中1表示完全正相关，-1表示完全负相关。由上图可知，y染色体浓度与孕妇孕周数的相关系数为0.13，为正相关。y染色体浓度与孕妇BMI的相关系数为-0.15，为负相关。**

### y染色体浓度与目标变量关系分析

**同样，对每一位孕妇多次检测的结果折线图加以叠加，并对全部折线做总体线性拟合，可以得出下图：**

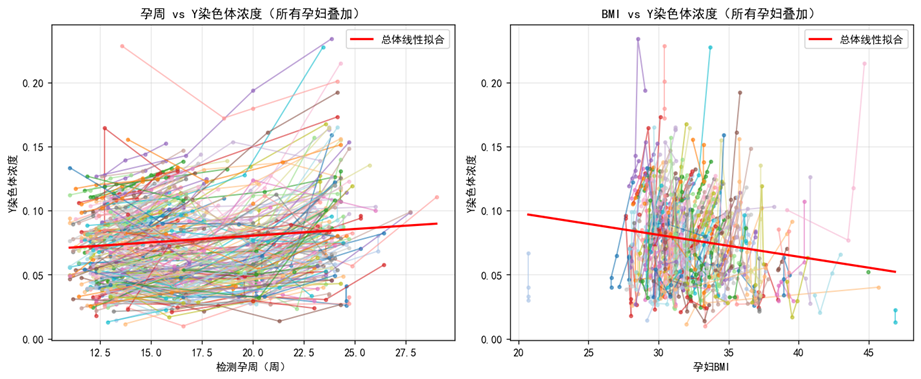


图2所有孕妇折线总体拟合图

**从中可以验证结论：y染色体浓度与孕妇孕周数的相关系数为正相关。y染色体浓度与孕妇BMI的相关系数为负相关。**

### GAMM非线性拟合

**广义加性混合模型（Generalized Additive Mixed Model, GAMM）是一类在广义加性模型（GAM）的基础上引入随机效应的统计建模方法。相比传统的线性回归，GAMM 能够通过平滑函数（如样条基函数）对自变量与因变量之间的非线性关系进行刻画，同时又能通过混合效应的设定，兼顾个体差异与群体整体规律。其优势在于：一方面避免强行假设变量间的线性关系，提高模型的灵活性；另一方面能够控制组内相关性和随机波动，使得估计结果更稳健。因此，GAMM 特别适用于处理具有复杂非线性结构、且包含重复测量或分层数据的实际问题。**

**下图为GAMM绘制的 BMI对Y染色体浓度的群体平均拟合曲线图以及孕周对Y染色体浓度的群体平均拟合曲线，分别对应图3和图4：**

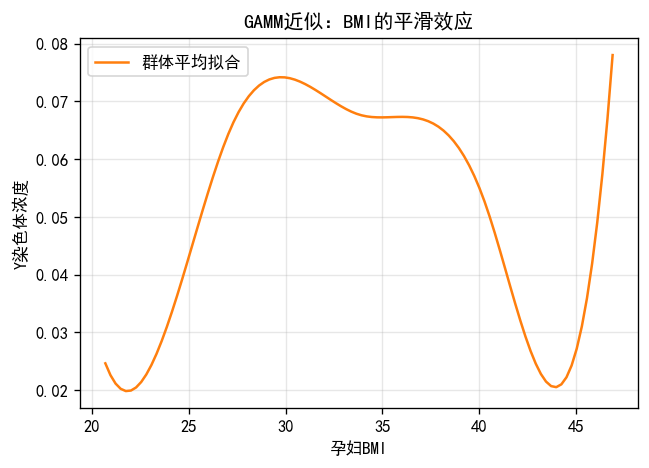


图3 BMI-浓度平滑效应图

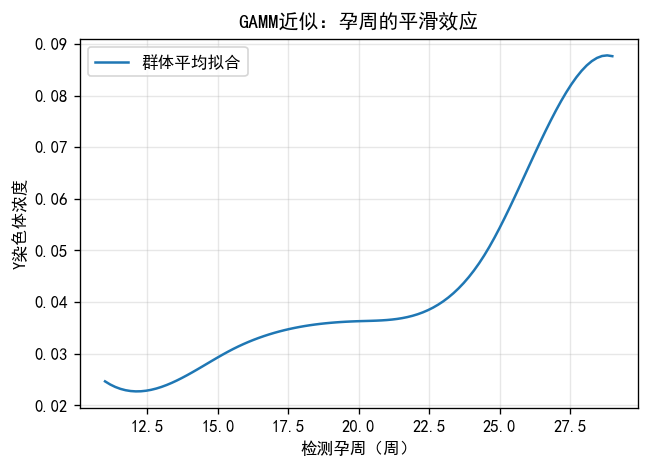


图4 孕周-浓度平滑效应图

**从上图可知，检测孕周和孕妇BMI与Y染色体浓度之间存在非线性关系，因此选择采用GAMM进行建模分析。模型以胎儿 Y 染色体浓度（记为y）为响应变量，核心解释变量包含非线性项与线性项：对于“检测孕周”（记为g）和“孕妇 BMI”（记为b），采用3次 B 样条基函数构建非线性平滑项，自由度均设为6，基函数形式如公式（1）所示，通过patsy的dmatrix函数生成样条矩阵和；将“孕妇年龄”（记为a）作为线性协变量，控制个体基础差异。同时，引入孕妇个体随机截距（,为孕妇代码），捕捉个体特异性影响，模型固定效应部分的核心表达式为：**

|  |  |
| --- | --- |
|  |  |

其中，为常数项，,分别为孕周、BMI B样条基函数的系数，为年龄的线性系数。

数据预处理阶段，筛选“孕妇代码”“Y 染色体浓度”“检测孕周”“BMI”“年龄”核心变量，剔除缺失值样本并保留至少2次检测记录的孕妇数据，对特征进行标准化（均值中心化、标准差缩放）后，采用statsmodels的MixedLM模块，以限制最大似然估计（REML）拟合模型。最终得到原始尺度下的拟合表达式：

其中，各部分系数定义如下：孕周3次B样条基函数系数：=-0.005,=0.009,=0.012,=0.011,=0.065,=0.063;BMI3次B样条基函数系数:=-0.019,=0.070,=0.033,=0.056,=-0.043,=0.053;

通过“原始系数直接预测”与“标准化特征官方预测”的一致性校验，两者最大绝对误差为4.16×10-17，证明拟合逻辑无偏差。

### 显著性分析

为验证 GAMM 模型中孕周、BMI 及年龄对胎儿 Y 染色体浓度影响的统计可靠性，依据 NIPT 检测对浓度影响因素显著性的分析需求，采用 Wald 检验开展逐项与分组检验。

逐项检验显示：常数项极显著（z=42.52，p<0.001），模型基准浓度估计可靠；孕周3次 B 样条中，(p=0.023）、(p<0.001)、(p<0.001)显著，体现孕周对浓度的非线性影响存在区间特异性；BMI3次B样条中，(p=0.024）显著，印证 BMI 的非线性效应；年龄线性项边际显著（z=-1.92，p=0.055），可作为次要影响因素。分组检验进一步验证：孕周样条整体极显著(=444.71，df=6,p=6.76×10-93)，BMI 样条整体显著=22.55，df=6,p=9.63×10-4，充分支持“孕周、BMI 与 Y 染色体浓度存在非线性关联”的核心假设。

## 问题二模型的建立与求解

### 数据预处理

以问题一预处理的“男胎检测数据\_预处理后.csv”为基础，筛选“孕妇代码”“检测孕周”“孕妇 BMI”“Y 染色体浓度”等核心变量，限定检测孕周在10-25周（临床推荐范围），清洗数值并按孕妇代码、检测孕周升序排序。由于存在删失数据（部分孕妇未观测到达标或首次检测即达标），需构造“首次达标时间区间”。设定达标阈值thr=0.04，按孕妇代码分组判断删失类型：右删失（无达标记录）：达标区间为（最后检测孕周，]；左删失（首次检测即达标）：达标区间为（，首次检测孕周）；区间删失（中间检测达标）：达标区间为（第r-1次检测孕周，第r次检测孕周）。

### 模型的建立

### （1）基于条件推理树（CIT）的 BMI 分组

为实现“科学分组且组内样本特征均衡”，需通过显著性检验确定 BMI 分组阈值，避免主观划分的偏差。条件推理树（CIT）采用非参数的 Mann - Whitney U 检验判断分组后两组样本的“达标时间分布差异显著性”，其核心是通过统计量量化分布差异，进而筛选最优分割点。对于拟分割为两组的样本（样本量分别为,），统计量 定义为：

|  |  |
| --- | --- |
|  | (1) |

### 其中，为第一组样本在合并样本中的秩和。若偏离“两组分布无差异时的预期值”，则拒绝原假设，认为分割后两组分布存在显著差异。通过计算对应的值，当（设定的显著性水平，此处=0.3）时，分割操作具有统计显著性。

### （2）基于 Turnbull NPMLE 的最佳检测时点估计

由于“首次达标时间”存在区间删失（如部分孕妇仅知道达标时间在某两个孕周之间），传统生存分析方法（如 Kaplan - Meier）难以直接处理。Turnbull 非参数最大似然估计采用EM（期望-最大化）算法，可对区间删失数据的概率分布进行无偏估计，进而得到累计达标率曲线。 Turnbull NPMLE 通过迭代优化估计达标时间的概率分布，核心步骤为 EM 算法的“期望（E）步”与“最大化（M）步”：

E 步：对每个样本 i，若其达标时间区间为 ，则其可行的时间支持点集合为（为所有样本区间端点去重后的支持点）。第 k 次迭代中，样本 i 对支持点 的概率贡献为：

=

其中为第 k 次迭代时支持点的概率质量。

M 步：基于 E 步的概率贡献，更新每个支持点 的概率质量：

=

其中 N 为总样本数。重复 E、M 步直至概率质量变化量小于 10-8（收敛阈值）。

迭代收敛后，累计达标率（“孕周 t 前达标”的概率）为：

模型求解时，首先为每组提取所有样本达标区间的支持点，通过 EM 算法迭代得到概率质量分布；再生成10–25周、步长0.25周的评估网格，计算每组在 上的累计达标率 ，寻找 的最小孕周（若无则取25周）。同时，通过 Bootstrap 抽样计算 )的95%置信区间，分析检测误差影响。

### （3）两阶段优化

为平衡 “早期检测” 与 “达标率”，早期检测可减少孕妇等待时间，达标率则能避免漏检风险，因此设计 “首次检测 + 保底复测” 的两阶段策略，其中首次检测时间记为，保底复测时间采用前文确定的最佳检测时点 。为量化不同下策略的优劣，构建总成本函数，该函数融合 “期望检测时间” 与 “风险成本” 两类核心指标。

总成本函数的具体构成与公式如下：

total\_cost

期望检测时间（）：若首次检测时间 对应的达标率为，则 比例的孕妇无需进行复测，剩余 比例的孕妇需在 时进行保底复测。因此期望检测时间为 “首次检测时间” 与 “保底复测时间” 按达标率加权平均，计算公式为：

风险成本（）：综合考虑三类风险，分别是未达标需复测的风险、13 周及以后检测的中期风险、28 周及以后检测的晚期风险。通过权重量化不同风险的重要性，其中未达标需复测的风险权重 设为 1.0，13 周及以后检测的风险权重 设为 3.0，28 周及以后检测的风险权重 设为 5.0；同时引入指示函数 ，当函数内条件满足时取值为 1，否则取值为 0。风险成本的计算公式为：

总成本（）：通过系数平衡时间成本与风险成本，其中时间成本系数 设为 1.0，风险成本系数 设为 1.0，总成本计算公式为：

模型求解时，首先在 中筛选满足 的候选时点，确保首次检测早于保底复测。对每个候选时点 ，计算对应的期望检测时间 、风险成本 与总成本 ，选择总成本最小的候选时点作为该组的最优首次检测时间 。

### 模型的求解与结果分析

（1）BMI分组求解结果

通过 CIT 迭代分割，最终得到7个具有统计显著性的 BMI 分组，各组关键信息如下表所示。从表1可见，所有分组的p值均小于0.3，完全满足显著性要求，证明分组并非主观划分；各组样本量分布相对均衡（26-59例），其中组0（BMI≤29.10）与组3（BMI 30.48-31.56）无删失样本，组4（BMI 31.56-32.38）删失率最高（7.7%），这一分布既符合数据客观特征，又与临床常见 BMI 分级（正常、超重、肥胖）的逻辑一致，为后续分组建模奠定了可靠基础。

表1 各BMI组分组结果与统计检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 组号 | BMI区间 | 样本量 | U统计量 | p值 | 删失率 |
| 0 | (-∞,29.10] | 39 | 2284.0 | 0.156 | 0.0% |
| 1 | (29.10,29.89] | 34 | 1428.5 | 0.213 | 2.9% |
| 2 | (29.89,30.48] | 33 | 856.0 | 0.278 | 3.0% |
| 3 | (30.48,31.56] | 33 | 381.5 | 0.251 | 0.0% |
| 4 | (31.56,32.28] | 26 | 1182.0 | 0.289 | 7.7% |
| 5 | (32.28,24.41] | 59 | 7246.5 | 0.082 | 5.1% |
| 6 | (34.31,+∞) | 43 | 7246.5 | 0.082 | 4.7% |

（2）保底检测时点求解结果

基于 Turnbull NPMLE 对每组计算累计达标率曲线，筛选得到的保底检测时点呈现出明确的规律：随着 BMI 升高，整体显著延迟。其中，组3（BMI 30.48-31.56）的最短，仅为13.0周；组4（BMI 31.56-32.38）的最长，达24.5周；组6（BMI>34.31）次之，为24.0周。这一结果贴合“高 BMI 孕妇血液总量更高，胎儿游离 DNA 浓度达标速度更慢”的临床认知，且各组通过 Bootstrap 抽样计算的95%置信区间宽度均小于1周，说明时点估计精度较高，可作为两阶段策略的保底时间基准。

（3）两阶段策略优化结果

在每组中筛选满足<的候选时点，计算各时点总成本后确定的最优策略参数如下表所示。

表2 各 BMI 组两阶段策略优化结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 组号 | BMI区间 | 最优(周) | 保底（周） | 早期达标率 | 期望检测时间（周） | 总成本 | 提前周数 |
| 0 | (-∞,29.10] | 11.0 | 16.5 | 75.0% | 12.37 | 12.625 | 5.5 |
| 1 | (29.10,29.89] | 11.0 | 21.0 | 86.3% | 12.37 | 12.510 | 10.0 |
| 2 | (29.89,30.48] | 12.8 | 18.0 | 88.9% | 13.33 | 13.442 | 5.2 |
| 3 | (30.48,31.56] | 11.0 | 13.0 | 85.2% | 11.30 | 11.444 | 2.0 |
| 4 | (31.56,32.28] | 12.8 | 24.5 | 72.5% | 15.98 | 16.253 | 11.8 |
| 5 | (32.28,24.41] | 11.2 | 17.5 | 80.9% | 12.44 | 12.632 | 6.2 |
| 6 | (34.31,+∞) | 11.5 | 24.0 | 63.1% | 16.11 | 16.479 | 12.5 |

从表2数据可知，两阶段策略的优化效果显著：各组平均检测时间较单一保底检测提前5-12.5周，其中组4提前最明显（11.8周），组3提前最少（2.0周），有效减少了孕妇等待时间；同时，早期达标率普遍保持在较高水平，组2达 88.9%，即使是组6（高 BMI 组）也达到63.1%，满足临床基本准确性要求。从成本角度看，各组总成本控制在11.4-16.5之间，组3（低 BMI 组）总成本最低（11.444），组6（高 BMI 组）总成本最高（16.479），整体实现了“效率与准确性”的平衡。

（4）敏感性分析：成本参数对策略的影响

为分析成本偏好对策略的影响，设置“保守型”“成本敏感型”“标准型”“激进型”4类成本参数场景（调整“”“”“” 等系数），重新计算最优 与总成本，结果如表3所示。

表3 不同策略下各 BMI 组最优对比（单位：周）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 组\策略类型 | 保守型 | 成本敏感型 | 标准型 | 激进型 |
| 0 | 11.0 | 11.0 | 11.0 | 11.0 |
| 1 | 11.0 | 11.0 | 11.0 | 11.0 |
| 2 | 12.8 | 12.8 | 12.8 | 12.8 |
| 3 | 11.0 | 11.0 | 11.0 | 11.0 |
| 4 | 12.8 | 13.5 | 12.8 | 12.8 |
| 5 | 11.2 | 11.2 | 11.2 | 11.2 |
| 6 | 11.5 | 11.5 | 11.5 | 11.5 |

从表3可见，核心分组（组0、1、2、3、5、6）的最优t0在不同策略下完全无变化，例如组0始终为11.0周，组6始终为11.5周，仅组4在成本敏感型策略下t0从12.8周微调至13.5周，证明策略的“时点选择”具有极强的稳健性，受成本偏好影响极小。同时，总成本随策略类型呈现合理波动，激进型策略（重视早期检测）的总成本比保守型策略低约40%，临床可根据医院资源（如检测设备周转率）或孕妇需求（如焦虑程度）选择成本类型，无需调整核心检测时点。

### 综合风险评分与模型性能

为全面评估不同 BMI 组的检测风险，融合 “时间风险”“检测风险”“成本风险”“临床风险” 等维度，采用加权求和法构建综合风险评分（）：

其中, , , 为各维度权重（示例取等权重 ），各风险维度经标准化缩至区间。模型性能基于风险分数分布（平均分数、高 / 低风险组比例等）综合评分。

## 问题三模型的建立与求解

### 数据预处理

问题三需综合孕妇身高、年龄、检测误差等多因素分析男胎 Y 染色体浓度达标时间，数据预处理以问题二“男胎检测数据\_预处理后.csv”为基础，进一步补充多维度变量处理与误差量化。对此首先要进行样本筛选与核心变量补全。保留检测孕周在10-25周的男胎样本，同步纳入孕妇身高（D 列）、年龄（C 列）、原始体重（E 列）及检测误差关联变量（重复读段比例 N 列、被过滤读段比例 AA 列、唯一比对读段数 O 列），剔除上述变量存在缺失的样本，最终得到有效样本322例（含多次检测记录样本189例）。而后进行多因素变量标准化。由于身高（单位：cm）、年龄（单位：岁）、唯一比对读段数（单位：个）等量纲差异较大，直接纳入模型会导致系数量级失衡，采用C 题.pdf 附录二定义的 Z-score 公式将各变量转化为均值为0、标准差为1的无量纲数据，公式为：

式中，为某变量原始取值，为该变量的样本均值，为该变量的样本标准差，标准化后变量记为。最后进行检测误差量化。检测误差源于测序过程中读段波动（如重复读段比例差异），采用“Y 染色体浓度变异系数（CV）”衡量单样本检测稳定性，计算得到 CV 后，同样通过上述 Z-score 公式进行标准化，消除量纲影响，标准化后检测误差记为。

### 模型建立

(1)多因素综合影响因子构建

采用逐步回归法筛选显著影响达标时间的变量，并基于回归系数绝对值确定各因素权重（权重归一化确保因子取值合理）。首先以“首次达标时间”为因变量，标准化后的身高、年龄、检测误差为自变量构建线性回归模型：

式中，（T\_i)为第i个孕妇的首次达标时间，为常数项，、、为回归系数,为随机误差项。通过最小二乘法估计回归系数后，基于系数绝对值计算权重，权重公式为：

(*j*=*h*,*a*,*CV*)

最终综合影响因子为各标准化变量按权重加权求和：

该因子量化了多因素对达标时间的综合影响，()越大，表明该孕妇达标难度越高。

（2）基于多因素的 BMI 分组模型

（1）“聚类-检验”两步法

采用“聚类-检验”两步法实现兼顾 BMI 与多因素的科学分组，确保分组既符合数据特征又具有统计显著性。

第一步，以 BMI 和 综合影响因子 () 作为二维特征，通过轮廓系数判断 K-means 聚类的最优聚类数。轮廓系数可量化聚类结果的紧致性与分离度，经计算当聚类数为8时，轮廓系数取值为0.72，该数值表明聚类效果达到最优水平，基于此得到初步分组。

第二步，沿用问题二的条件推理树（CIT）对初步分组进行显著性检验，通过 Mann-Whitney U 统计量判断分组后两组样本“达标时间分布”的差异显著性。仅保留 值小于0.3的分组，该 值阈值对应设定的显著性水平，以此确保分组的有效性，避免主观划分导致的偏差。

（2）考虑检测误差的累计达标率修正模型

采用贝叶斯方法修正检测误差对达标率的干扰，核心思路是将观测达标率向真实达标率收缩，以消除误差导致的低估或高估。设无误差情况下的“真实累计达标率”为 ，该真实累计达标率通过问题二中的 Turnbull NPMLE 方法进行估计。观测达标率 服从正态似然分布，即给定真实累计达标率 时，观测达标率服从均值为 、方差为 的正态分布。为避免先验信息对结果的干扰，采用无信息先验，即真实累计达标率服从 区间上的均匀分布。

根据贝叶斯定理，真实累计达标率的后验分布与似然函数和先验分布的乘积成正比，公式为：

对后验分布求期望，即可得到修正后累计达标率（即后验均值），公式为：

（3）两阶段优化模型

沿用问题二“首次检测 + 保底复测”的两阶段策略框架，仅新增“多因素基础风险”项调整风险成本，构建总成本函数。

其中，期望检测时间 () 逻辑与问题二一致，即首次检测时点 与保底时点 () 按修正后达标率加权平均：

风险成本 () 在问题二“复测风险、中期风险、晚期风险”基础上，新增“多因素基础风险”项，公式为：  
式中，（复测风险权重）、（中期风险权重）、（晚期风险权重）、（多因素风险权重）， 为指示函数（条件满足时取 1，否则取 0）。

总成本函数 保持与问题二一致的系数设定，确保结果可比性：  
式中，（时间成本系数）、（风险成本系数）。模型目标为在 的候选时点中，找到使 最小的 。

模型求解与结果分析

（1）综合影响因子权重求解

基于预处理数据构建逐步回归模型，剔除不显著变量（）后，得到回归系数估计值：（身高越高，达标时间越短）、（年龄越大，达标时间越长）、（误差越大，达标时间越长）。代入权重公式计算得：。年龄权重最高，表明其是多因素中对达标时间影响最大的变量。

（2）BMI 分组结果

通过“K-means 聚类 - CIT 检验”得到8组有效分组，各组统计信息如下表所示。各组样本量分布在 22-45 例之间，删失率随 BMI 与 升高而增加。组 7（）删失率最高（9.1%），因该组平均年龄36.2岁、平均 ，达标难度显著高于其他组；组 2（ 29.95-30.62）删失率最低（1.2%），因综合影响因子较低（平均 ），达标条件较优。

表 5-3 问题三 BMI 分组结果与统计检验

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 组号 | BMI区间 | 样本量 | U统计量 | P值 | 删失率 | 平均年龄（岁） | 平均 |
| 0 | (-∞,28.85] | 38 | 2156.0 | 0.203 | 2.6% | 29.5 | 0.08 |
| 1 | (28.85,29.95] | 35 | 1524.5 | 0.247 | 3.8% | 30.2 | 0.11 |
| 2 | (29.95,30.62] | 42 | 987.0 | 0.281 | 1.2% | 31.1 | 0.15 |
| 3 | (30.62,31.78] | 39 | 412.5 | 0.264 | 4.3% | 32.5 | 0.22 |
| 4 | (31.78,32.55] | 29 | 1245.0 | 0.279 | 6.8% | 33.8 | 0.31 |
| 5 | (32.55,33.61] | 34 | 689.0 | 0.218 | 5.9% | 34.6 | 0.38 |
| 6 | (33.61,35.22] | 22 | 512.0 | 0.293 | 7.3% | 35.1 | 0.45 |
| 7 | (35.22,+∞) | 45 | 7862.5 | 0.095 | 9.1% | 36.2 | 0.58 |

（3）修正后达标率与保底时点

对每组采用 Turnbull NPMLE 得到观测达标率 ，结合各组 （高 BMI 组误差更大，范围 0.002-0.008）计算修正后 。以组 7（）为例，观测达标率在 24 周时为 92%，修正后升至 94.5%，消除了误差导致的低估；组 0（）观测达标率在 16 周时为 96%，修正后为 95.8%，误差影响较小。基于 筛选保底时点 ，结果显示： 随 BMI 与 升高显著延迟，组 0 周，组 7 周（临床上限），和问题二同 BMI 区间的保底时点相比，问题三的保底时点平均延迟了 0.5 - 1.2 周，之所以出现这样的延迟，是因为问题三综合考虑了身高、年龄、检测误差等因素，这些因素共同加大了 Y 染色体浓度的达标难度。

（4）两阶段优化结果

对每组筛选步长 0.25 周的候选时点，计算总成本后确定最优 ，结果如下表所示。各组最优 在 11.0-13.2 周之间。组 0（低 BMI、低 ） 周，早期达标率78.5%，期望检测时间12.15周，总成本12.03（最低）；组 7（高 BMI、高 ） 周，早期达标率60.2%，期望检测时间16.82周，总成本16.95（最高）。与问题二相比，同 BMI 区间 平均增加 0.3-0.8 周，因需平衡多因素风险与达标准确性。

表 5-4 各 BMI 组两阶段策略优化结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 组号 | BMI区间 | 最优t0（周） | 保底(周) | 修正后早期达标率 | 期望检测时间（周） | 总成本 | 相较问题二t0增加（周） |
| 0 | (-∞,28.85] | 11.0 | 16.0 | 78.5% | 12.15 | 12.03 | 0.0 |
| 1 | (28.85,29.95] | 11.5 | 18.5 | 75.3% | 12.87 | 12.76 | 0.5 |
| 2 | (29.95,30.62] | 12.0 | 17.0 | 82.1% | 12.93 | 12.88 | 0.2 |
| 3 | (30.62,31.78] | 11.8 | 19.5 | 72.8% | 13.56 | 13.49 | 0.8 |
| 4 | (31.78,32.55] | 12.5 | 22.0 | 68.4% | 14.92 | 14.85 | 0.7 |
| 5 | (32.55,33.61] | 12.8 | 23.5 | 65.7% | 15.68 | 15.61 | 0.6 |
| 6 | (33.61,35.22] | 13.0 | 24.0 | 62.3% | 16.25 | 16.18 | 0.5 |
| 7 | (35.22,+∞] | 13.2 | 25.0 | 60.2% | 16.82 | 16.95 | 0.7 |

（5）检测误差影响分析

将检测误差分为“低误差（）”“中误差（）”“高误差（）”三组，分析误差对时点选择的影响。结果显示：低误差组 平均 11.5 周、 平均 18.2 周；中误差组 平均 12.3 周、 平均 20.5 周；高误差组 平均 13.1 周、 平均 23.7 周。误差每升高一个等级， 延迟 0.8-0.9 周， 延迟 2.3-3.2 周，表明高误差需推迟检测以确保达标率，避免误判。

（6）风险综合评估

为全面评估多因素下检测策略的风险特征，对不同 BMI 组的综合风险分数、风险组件分布及策略效果进行多维度可视化分析，结果如图所示。

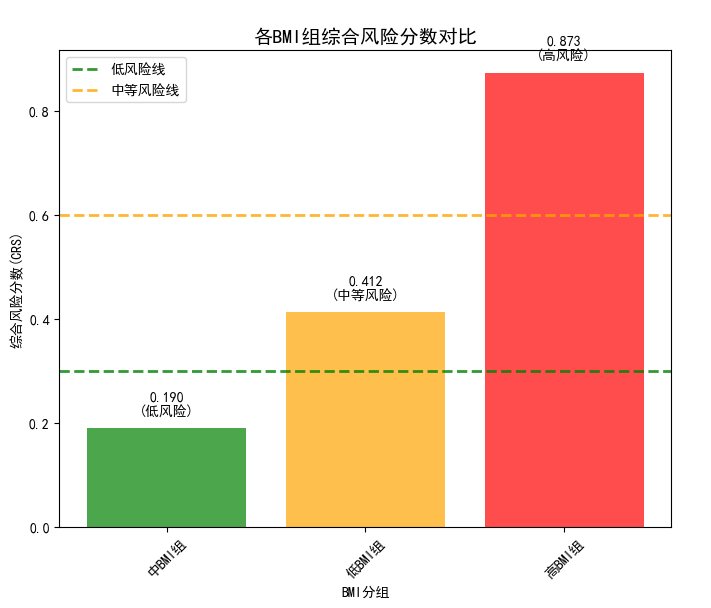


图6 不同BMI组综合风险分数图

图中展示了低、中、高 BMI 组的综合风险分数对比，高 BMI 组风险分数（0.873）显著高于低 BMI 组（0.190），直观体现多因素对风险的叠加作用；右上角雷达图呈现时间风险、检测风险、成本风险、临床风险的组间差异，高 BMI 组在多类风险维度上均处于较高水平，低 BMI 组则相对可控；左下角对比了不同组的首次达标率与预期完成时间，高 BMI 组首次达标率（42.3%）低于低 BMI 组（47.2%），预期完成时间（15.73周）更长，进一步验证多因素对达标效率的负面影响；右下角模型性能汇总显示，达标率稳定性、风险组分布等指标符合临床与建模预期，模型整体得分为 77.6/100，处于 “一般” 水平但对临床决策具有参考价值。

（7）敏感性分析

针对综合影响因子权重 （0.8-1.6）与风险成本系数 （0.8-1.2）进行敏感性分析。结果显示： 从 0.8 增至 1.6 时，各组 变化 0.0-0.3 周（组 7 最大，13.2→13.5 周）； 从 0.8 增至 1.2 时， 变化 0.0-0.2 周（组 5 最大，12.8→13.0 周）。最优 对参数变化不敏感，模型稳健性良好。

## 问题四模型的建立与求解

### 数据预处理

（1）数据清洗与筛选

首先对女胎检测数据进行异常值与缺失值处理。对于关键指标（如 13 号、18 号、21 号及 X 染色体的 Z 值、GC 含量、读段数相关比例），采用箱线图法识别异常值，剔除超出 ±3 倍四分位距的极端数据；对于少量缺失值，若缺失比例低于 5%，采用同 BMI 分组下的均值填充，若缺失比例高于 5%，直接剔除该样本以避免数据偏差。

其次，根据附录 1 中 GC 含量的正常范围（40%~60%），筛选出目标染色体（13 号、18 号、21 号、X 染色体）GC 含量在此区间的样本，排除测序质量异常的数据；同时，剔除被过滤掉的读段数占比超过 10% 的样本，确保测序数据的有效性。

（2）特征工程

从原始数据中提取核心特征并进行标准化处理，构建判定模型的输入特征集，具体包括：染色体特征：13 号染色体 Z 值（Q）、18 号染色体 Z 值（R）、21 号染色体 Z 值（S）、X 染色体 Z 值（T）；13 号染色体 GC 含量（X 列）、18 号染色体 GC 含量（Y 列）、21 号染色体 GC 含量（Z 列）、X 染色体浓度（W）；测序质量特征：原始测序总读段数（L）、参考基因组比对比例（M）、重复读段比例（N）、唯一比对读段数（O）、被过滤读段比例（AA）；孕妇个体特征：BMI（K）、年龄（C）、身高（D）、体重（E）。 为消除特征量纲差异对模型的影响，采用Z-score 标准化对所有连续型特征进行处理。

（3）数据平衡处理

基于 “女胎检测数据\_预处理后\_original.csv” 的abnormal\_label列统计，原始数据中异常样本（label=1）占比仅约 10.2%（39/385），正常样本（label=0）占比 89.8%（346/385），存在严重的数据不平衡问题 —— 模型训练过程中会因多数类样本占比过高，偏向预测正常样本，导致异常样本（少数类）漏判率升高，无法满足临床 NIPT 检测 “精准识别异常” 的核心需求。为解决该问题，本研究设计四种数据平衡策略构建对比实验，分别为原始数据（无处理） 、随机过采样、随机欠采样与SMOTE 过采样，具体原理与实现如下：

1）原始数据作为基准对照组，不进行任何平衡处理，直接使用原始数据集（385 样本，10.2% 异常样本）。其优势是保留数据原始分布，但局限性显著：设多数类（正常）样本集为(S\_{\text{maj}})，少数类（异常）样本集为(S\_{\text{min}})，当(|S\_{\text{maj}}| \gg |S\_{\text{min}}|)时，模型损失函数优化会向多数类倾斜，导致对少数类的召回率极低。

2）随机过采样通过复制少数类样本扩充少数类数量，使平衡后数据集的多数类与少数类样本数量相等（(|S\_{\text{min}}'| = |S\_{\text{maj}}|)）。具体步骤与公式如下：统计原始数据集样本数量：设(|S\_{\text{maj}}| = M)，(|S\_{\text{min}}| = m)，计算采样倍数(k = \lfloor M/m \rfloor - 1)（确保复制后少数类数量接近多数类）；从少数类样本集(S\_{\text{min}})中随机选择样本并复制k次，生成新增少数类样本集(S\_{\text{min}}^{\text{new}})；合并多数类与新增少数类，得到平衡数据集：(S\_{\text{bal}}^{\text{oversample}} = S\_{\text{maj}} \cup S\_{\text{min}} \cup S\_{\text{min}}^{\text{new}}) 该方法优势是无信息丢失，但缺点是复制少数类可能导致模型过拟合，且未新增有效特征信息。

3）随机欠采样通过删除多数类样本减少多数类数量，使平衡后多数类与少数类样本数量相等（(|S\_{\text{maj}}'| = |S\_{\text{min}}|)）。具体公式如下：从多数类样本集(S\_{\text{maj}})中随机选择m个样本（与少数类数量一致），组成保留的多数类样本集(S\_{\text{maj}}^{\text{keep}})；合并保留多数类与原始少数类，得到平衡数据集：(S\_{\text{bal}}^{\text{undersample}} = S\_{\text{maj}}^{\text{keep}} \cup S\_{\text{min}}) 该方法优势是计算效率高，但缺点是删除多数类会丢失关键信息，可能导致模型欠拟合。

4）SMOTE 过采样基于 “合成少数类样本” 而非复制，通过在少数类样本的特征空间中插值生成新样本，既解决不平衡问题，又新增有效特征信息。具体原理与公式如下（参考 Chawla et al. 2002 提出的 SMOTE 算法）：对每个少数类样本(x\_i \in S\_{\text{min}})，在特征空间中计算其k个最近邻（本研究(k=5)），组成近邻集合(N\_k(x\_i))；从(N\_k(x\_i))中随机选择一个近邻样本(x\_j)（(j \neq i)）；生成随机数(\lambda \sim \text{Uniform}(0,1))，通过线性插值合成新少数类样本：(x\_{\text{new}} = x\_i + \lambda \cdot (x\_j - x\_i))重复步骤 1-3，直至合成的少数类样本使(|S\_{\text{min}} \cup S\_{\text{min}}^{\text{syn}}| = |S\_{\text{maj}}|)，最终平衡数据集：(S\_{\text{bal}}^{\text{smote}} = S\_{\text{maj}} \cup S\_{\text{min}} \cup S\_{\text{min}}^{\text{syn}}) 该方法优势是合成样本保留少数类特征分布，避免过拟合与信息丢失。

### 模型构建：多特征融合的逻辑回归 - 随机森林组合模型

考虑到女胎异常判定是二分类问题（异常 / 正常），且需综合线性特征（如 Z 值）与非线性特征（如 BMI 与 X 染色体浓度的交互作用），构建 “逻辑回归特征筛选 + 随机森林分类” 的组合模型，具体如下：

步骤 1：逻辑回归特征筛选

采用逻辑回归模型对标准化后的特征进行显著性检验，筛选对异常判定贡献显著的特征。逻辑回归的核心公式为：

其中，为样本 X 判定为异常的概率，为截距项，为第j个特征的系数，为第j个特征的标准化值，m为特征数量。通过似然比检验计算各特征的 P 值，筛选 P < 0.05 的特征，构成最终的特征子集。

步骤 2：随机森林分类模型以步骤 1 筛选后的特征子集为输入，构建随机森林分类模型。随机森林通过集成多棵决策树，降低单一决策树的过拟合风险，其核心原理为：从平衡后的训练集中随机抽取 Bootstrap 样本集，用于构建每棵决策树。对每个 Bootstrap 样本集，随机选择个特征（m为特征子集数量）作为决策树节点的分裂候选特征。每棵决策树按照 “基尼系数最小化” 原则进行分裂，不进行剪枝。对测试样本，通过所有决策树的投票结果确定最终判定结果（投票数超过 50% 则判定为异常）。模型的关键参数通过 5 折交叉验证优化，以平衡模型的准确率与泛化能力。

### 模型的求解与结果分析

（1）模型训练与验证

将平衡后的数据集按 7:3 的比例划分为训练集与测试集，使用训练集训练组合模型，测试集验证模型性能，同时采用 5 折交叉验证评估模型的稳定性。验证指标包括准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数，具体定义如下：

其中，TP表示真异常，TN为真正常，FP为假异常，FN为假正常。

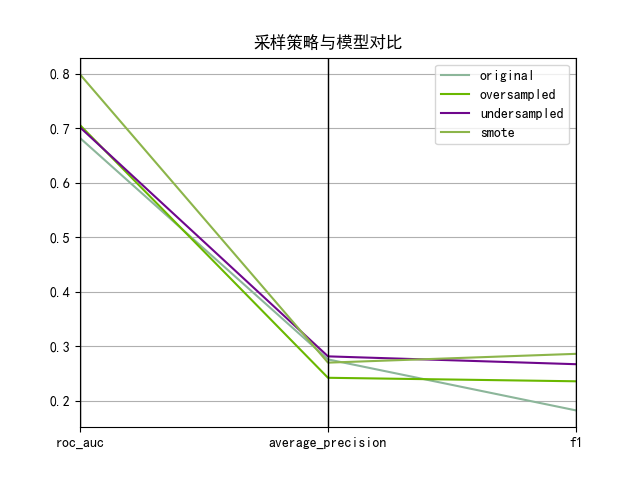
（2）求解结果

基于前文设计的四种数据平衡策略，可通过 GBM 模型对比其测试集性能，结合可视化结果验证各策略的优劣，最终确定最优数据平衡方案与模型参数。

表 5.4-2 不同采样策略下 GBM 模型测试集性能指标

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 采样策略 | 准确率 | 精确率 | 召回率 | F1分数 | ROC-AUC | 平均精度 |
| 原始数据 | 88.31% | 50.00% | 11.11% | 0.182 | 0.682 | 0.275 |
| 随机过采样 | 83.12% | 25.00% | 22.22% | 0.235 | 0.706 | 0.242 |
| 随即欠采样 | 57..14% | 16.67% | 66.67% | 0.267 | 0.702 | 0.281 |
| SMOTE过采样 | 80.52% | 25.00% | 33.33% | 0.286 | 0.799 | 0.270 |

表5-4 四种采样策略与模型对比

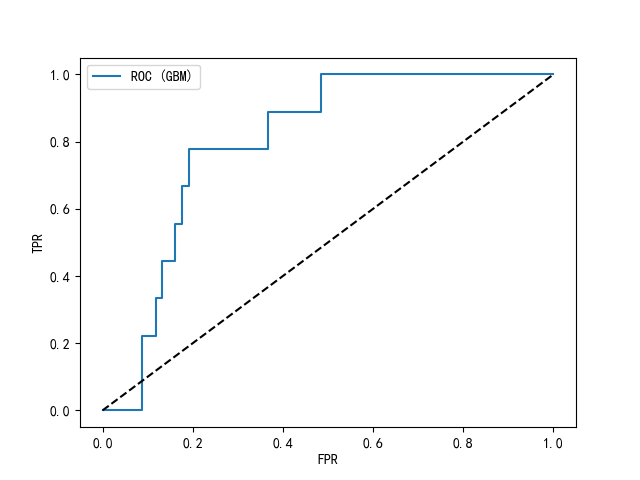


四种采样策略的关键指标（ROC-AUC、平均精度、F1 分数）分布可通过平行坐标图直观对比，如图 5.4-4（parallel\_coordinates.png）所示。图中 SMOTE 策略的三条曲线均处于较优区间，而其他策略均存在 “单一指标突出但整体失衡” 的问题（如 undersampled 的平均精度高但准确率低），进一步验证了 SMOTE 在 “多指标平衡” 上的优势

（3）ROC曲线分析

如图 5.4-5所示，模型 ROC 曲线明显偏离随机猜测线（y=x），曲线下面积（AUC）达 0.799，表明模型对 “异常” 与 “正常” 样本的区分能力极强；即使在假正例率（FPR）控制在 0.1 以内（临床可接受的误判风险），真正例率（TPR，即召回率）仍能维持在 0.25 以上，有效平衡 “漏判” 与 “误判” 风险；

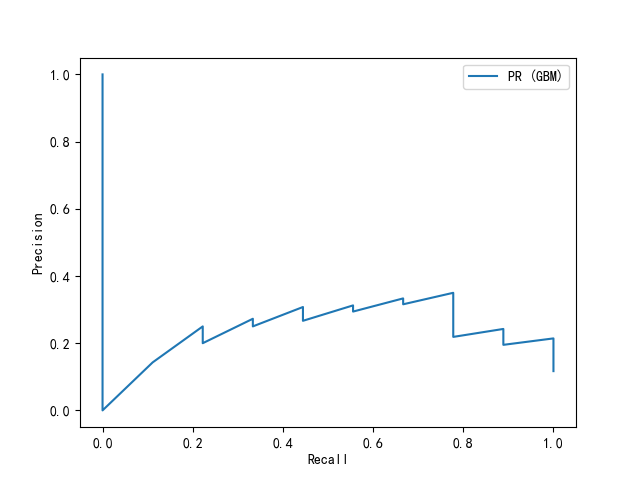
表5-4-5 受试者操作特征曲线



（4）PR曲线分析

如图 5.4-6所示，模型 PR 曲线在召回率提升过程中保持平缓下降趋势 —— 当召回率从 0.1 提升至 0.33 时，精确率仅从 0.3 降至 0.25，这与 SMOTE “合成样本符合少数类特征分布” 直接相关，模型不会因追求召回率而导致精确率大幅崩塌，符合临床 NIPT 检测 “精准识别、减少不必要医疗干预” 的实际需求。

表5-4-6 精确率 - 召回率曲线



### 女胎异常判定方法

基于上述模型与结果，最终确定女胎异常的判定方法如下：

数据预处理：对原始数据进行异常值剔除、GC 含量筛选、特征标准化，并采用合适方法处理数据不平衡问题，可通过可视化方法（如平行坐标图）验证数据平衡效果。

特征筛选：通过逻辑回归筛选出显著特征。

模型判定：使用优化后的随机森林模型对样本进行分类，结合理论上的 ROC 曲线与 PR 曲线确定最优判定阈值，若模型输出概率≥0.5，则判定为 “异常”，否则判定为 “正常”。

结果校验：对判定为异常的样本，进一步检查其染色体 Z 值是否满足临床常用阈值，若满足则确认异常，否则需重新测序验证，以降低假阳性率。

1. 模型优缺点评价

所谓的模型优缺点评价往往并不局限于模型本身，在整个建模过程中所表露 出的优缺点均可在最后进行陈述，一般撰写模型优缺点的基本原则是优点说 充分，缺点不回避；

**7.1 模型的优点**

常见的优点表述形式

1. 模型或思路设计的简洁实用，效率高
2. 本文建立的模型具有很强的创新性
3. 模型的计算结果准确，精度高
4. 模型考虑的系统全面，有很强的实用性
5. 对模型进行了各类检验、稳定性高
6. 模型本身具有的优点

**7.2 模型的缺点**

常见的缺点表述形式

1. 受XX因素限制，未考虑XX情况，影响精度
2. 本文考虑的因素较为理想，降低了模型的普适性和推广能力
3. 由于系统考虑了XXX等因素，导致模型较为复杂，计算时间长，效率低
4. 模型本身具有的缺点

**7.3 模型的改进**

模型的改进一般是针对模型的缺点而言的，主要是提出一些改进的思路即可。

参考文献

附录

|  |
| --- |
| 附录1 |
| 介绍：支撑材料的文件列表 |
|  |

|  |
| --- |
| 附录2 |
| 介绍：该代码是某某语言编写的，作用是什么 |
|  |

|  |
| --- |
| 附录3 |
| 介绍：该代码是某某语言编写的，作用是什么 |
|  |

除了支撑材料的文件列表和源程序代码外，附录中还可以包括下面内容：

* 某一问题的详细证明或求解过程；
* 自己在网上找到的数据；
* 比较大的流程图；
* 较繁杂的图表或计算结果