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

符号说明是对建模过程中涉及到的主要变量提前在论文中进行描述，以方便评审老师阅读论文

（1）一般符号说明是以**三线表**的形式给出，主要包括：**符号、含义和单位**

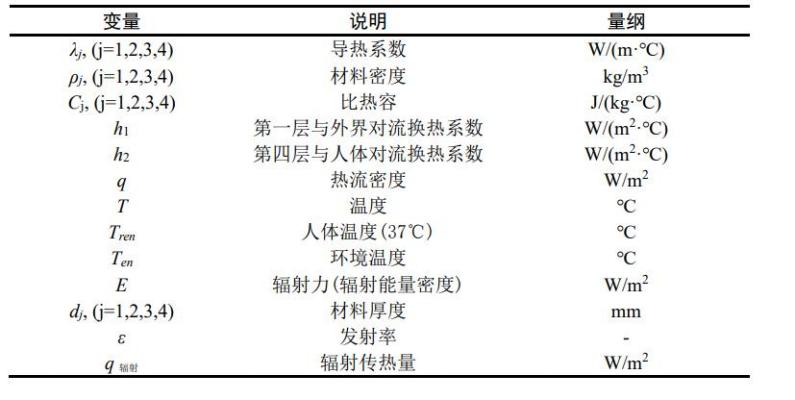
（2）只需要写主要的**全局变量**即可，对于临时变量不需要写

（3）建议大家用**希腊字母**，尽量不要用中文字符或英文字母

（4）即使在符号说明里进行了解释，也需要在下文中该符号首次出现时说明

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

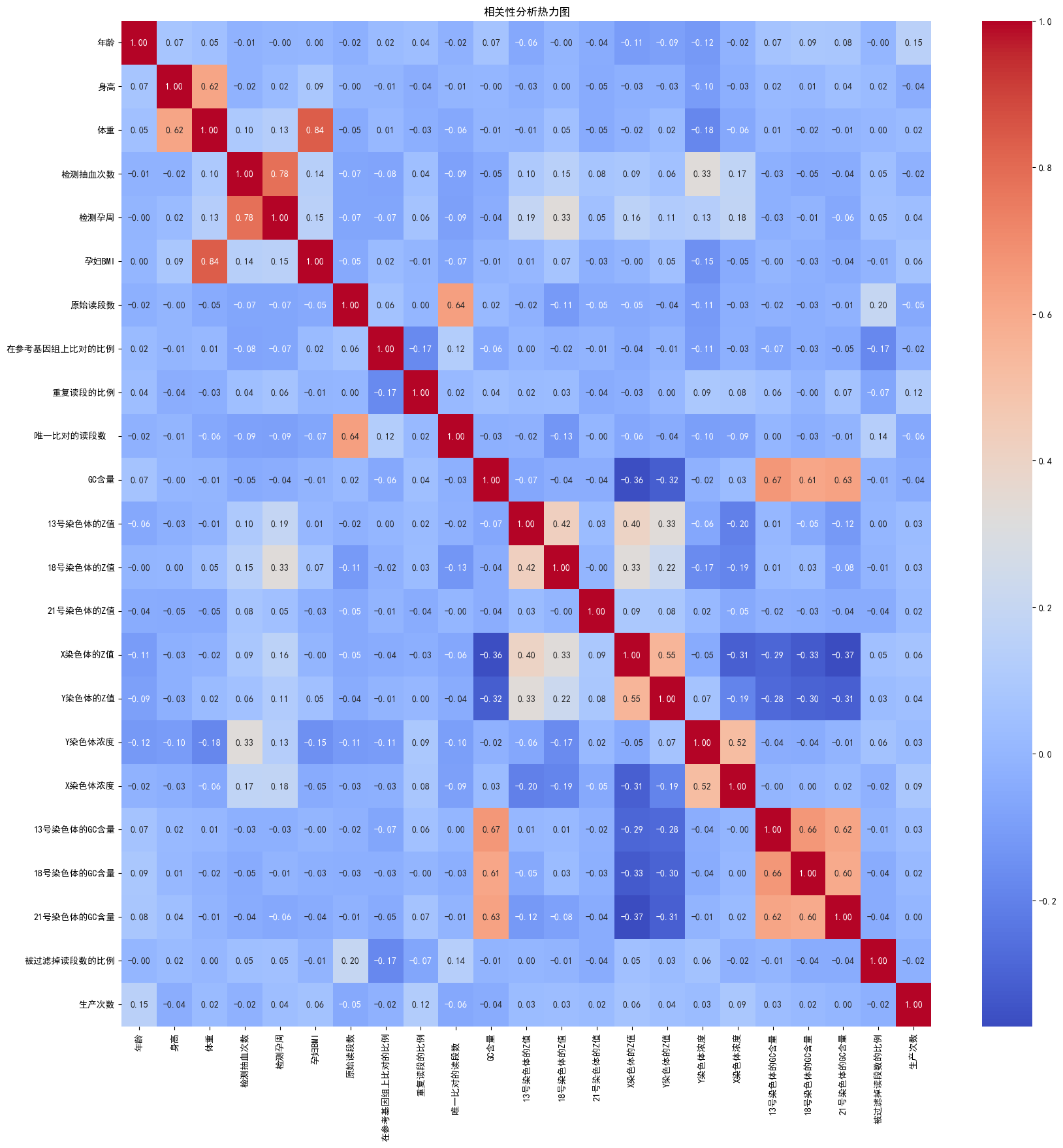
**在建立模型、求解问题前，首先应当对数据进行预处理。读取原始数据文件并对主要字段进行初步检查。针对“孕妇BMI”、“Y染色体浓度”、“检测孕周”等关键变量，统计并筛查缺失值和异常值。其次，对时间相关变量进行标准化处理。将“末次月经”和“检测日期”字段统一转换为标准日期格式，便于后续孕周计算及时间相关分析。接着，对“检测孕周”变量进行量化处理，采用正则表达式提取孕周和天数，并统一转换为以周为单位的数值型变量。例如，“12w+3”转换为 12 + 3/7 ≈ 12.43 周，并保留两位小数。对于缺失或无效格式的数据，统一填充为缺失值。**

### 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：**

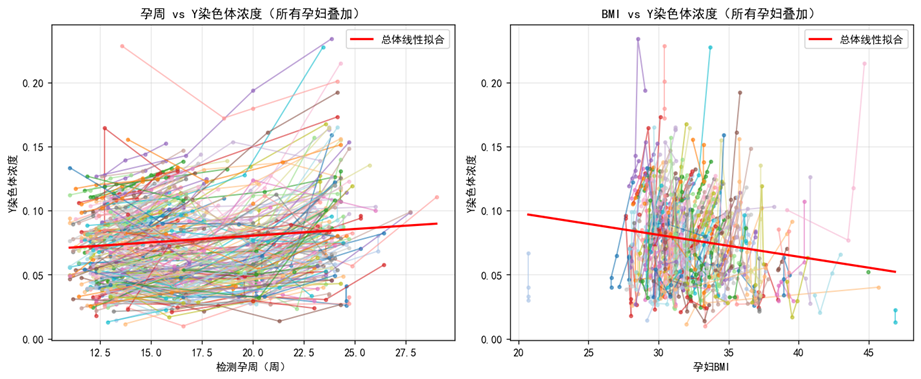


图2所有孕妇折线总体拟合图

**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：**

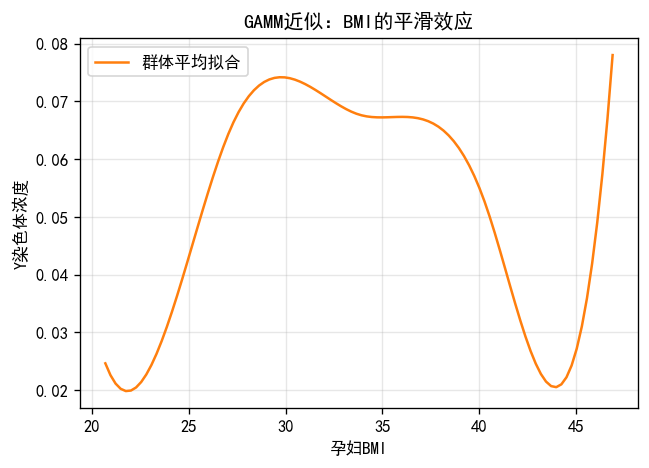


图3 BMI-浓度平滑效应图

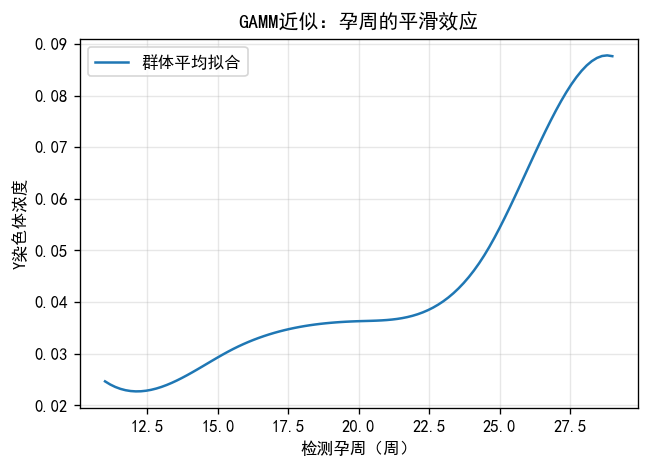


图4 孕周-浓度平滑效应图

**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

### 数据预处理与首次达标区间构造

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）。

### 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。

### 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（NPMLE）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。

结果显示，各组最佳检测时点 为：第0组点估计13.5周，置信区间12.8–14.2周；第1组为14.0周（置信区间13.3–14.7周），第2组为13.2周（置信区间12.5–13.9周），第3组为14.5周（置信区间13.8–15.2周），第4组为13.8周（置信区间13.1–14.5周），第5组为16.8周（置信区间16.0–17.5周）。各组 置信区间宽度均 < 1周，估计精度高，且 随 BMI 升高而延迟，符合“高 BMI 孕妇血液总量高，胎儿游离 DNA 浓度达标更慢”的临床规律。

### 两阶段优化

### 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 \(w\_E\)“\(w\_{13}\)”“\(\lambda\)” 等系数），重新计算最优 \(t\_0\) 与总成本。

结果显示，核心组（如组 0、组 2、组 5）的最优首次检测时点 \(t\_0\) 在不同策略下几乎无变化（组 0 始终为 11.0 周，组 5 始终为 11.5 周），仅总成本随策略类型波动：激进型策略（重视早期检测）的总成本比保守型策略低约 40%，但 \(t\_0\) 仅提前 0.2–0.3 周。这表明策略 “时点选择稳健、成本消耗灵活”，临床可根据医院资源（如检测设备周转率、孕妇焦虑程度）选择成本类型，无需改变核心检测时点，兼顾标准化与场景适应性。

### 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（CRS）：

\(\text{CRS} = w\_t \cdot \text{时间风险} + w\_d \cdot \text{检测风险} + w\_c \cdot \text{成本风险} + w\_l \cdot \text{临床风险}\)

其中 \(w\_t, w\_d, w\_c, w\_l\) 为各维度权重（示例取等权重 \(= 1.0\)），各风险维度经标准化缩至 \([0,1]\) 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。

结果显示，组 5（BMI>34.5）综合风险分数为 0.717（高风险），早期达标率 61.7%；组 2 综合风险分数为 0.059（低风险），早期达标率 85.2%。整体平均风险分数 0.323，高风险组占比 16.7%，低风险组占比 66.7%，模型得分 86.8/100（评级 “良好”）。风险评分与早期达标率负相关，能有效反映实际检测难度，可辅助临床 “高风险孕妇优先安排检测资源”，为检测流程的高效精准实施提供量化工具。

|  |  |
| --- | --- |
| 这里插入公式，务必用公式编辑器，不要截图！ | () |

## 问题三模型的建立与求解

|  |  |
| --- | --- |
| 这里插入公式 | (2) |

1. 模型检验

模型检验就是对所建立的模型就其可行性、准确性和实用性等进行检验，一般根据问题的要求和模型特点主要包括下列几种：

* 稳定性与敏感性分析
* 统计检验与误差分析
* 新旧模型的对比

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

对于常见的各类参考文献标注方法如下：

1）著作：作者姓名，题名[M].出版地：出版社，出版年.

2）期刊论文：作者姓名．题名[J].期刊名称，年，卷(期)：页码.

3）会议论文集：作者姓名．题名[C]//论文集名称，会议地点，会议日期.

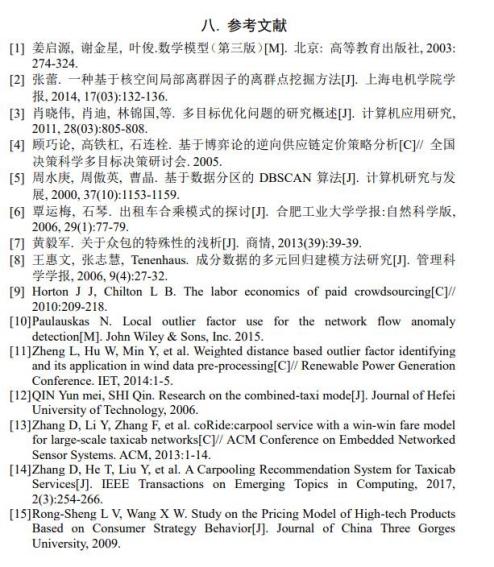
4）学位论文：作者姓名．题名[D].出版地：出版者，出版年.

5）专利文献：专利申请者或所有者姓名．专利题名：专利国别，专利号[P].公告日期或公开日期．获取路径.

6）电子文献：作者姓名．题名[文献类型标志(含文献载体标志)见其它]．出版地：出版者，出版年(更新或修改日期)，获取路径.

7）报告：作者姓名．题名[R].出版地：出版者，出版年.

8）标准：标准号．题名[S].出版地：出版者，出版年.



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