基于逻辑回归与生存分析的 NIPT 检测时点优化及异常判定研究

摘要

为应对无创产前检测（NIPT）在临床应用中面临的“一刀切”检测方案导致效率与准确性不足的挑战，尤其是在高身体质量指数（BMI）孕妇群体中，本文旨在构建一个**多阶段、个体化的 NIPT 检测决策支持与异常判定模型框架**。研究致力于通过**系统性建模**，优化检测时点选择，从而**最小化潜在临床风险**，并为女胎染色体异常提供一个高效的判别方法。

针对问题一: 本研究的亮点在于构建了一个**非线性混合效应模型**，创新性地将孕妇个体差异作为随机效应处理。研究发现，**孕妇个体差异**是解释浓度变异的最主要因素，贡献了**约 66%** 的变异，远超孕周与 BMI 的固定效应。该模型精准捕捉了浓度与 BMI 的非线性关系，相较于传统回归模型，在解释度（**条件 R^2 高达 0.7888**）与模型择优准则（**AIC 值更低，为-4392.56**）上均表现出显著优势，为后续个体化决策奠定了坚实基础。

针对问题二：本文创新性地引入**生存分析框架**来科学处理“事件是否发生及其发生时间”的**右删失数据**。通过**基于对数秩检验的递归分割算法**，将孕妇客观地划分为四个其达标时间分布存在显著差异的 BMI 组别。在此基础上，构建了一个权衡“检测失败风险”与“诊疗延迟风险”的**综合损失函数**，求解得到各组的最优 NIPT 时点（例如，对于 BMI 最高的 **[35.06, 46.88]** 组别，最优检测时点为 **19.43 周**），实现了从静态分组到动态优化的方法学突破。

针对问题三：为进一步提升决策的精准性，研究深化至多维因素分析。对于各 BMI 分组，分别构建**多变量逻辑回归模型**，其创新点在于揭示了不同 BMI 群体中**主导影响因素的异质性**（例如，低 BMI 组中**体重**影响显著，而高 BMI 组中**年龄**则呈现负向影响）。据此将最优检测时点进行了精细化调整，例如将部分高 BMI 组的最佳时点**提前至 10 周**。蒙特卡洛模拟进行的稳健性分析量化了检测误差的影响，证实极高 BMI 群体对检测精度要求更高（当误差标准差为 0.015 时，**假阴性率 FNR 飙升至 10.06%**）。

针对问题四：本研究以临床应用为导向，构建了一个基于**机器学习**的二分类判别模型。模型选择的亮点在于，不以整体准确率为唯一标准，而是优先考虑**异常样本召回率**，最终选择了在该指标上表现最优（**高达 77.78%**）的**逻辑回归模型**，以最大限度规避临床漏诊风险。研究还发现，衍生的**“测序质量”**综合指标是判别的最核心特征，并基于 ROC 曲线确定了 **0.4** 作为兼顾漏诊与误诊风险的最优判定阈值。

关键词: **无创产前检测 (NIPT)**；**混合效应模型**；**生存分析**；逻辑回归；决策优化

一、 问题重述

1.1 问题背景

随着我国社会经济的发展和民众生活质量的提高，优生优育已成为重要的家庭与社会关注点。《“健康中国 2030”规划纲要》指出我国正积极构建和完善包括产前筛查与诊断在内的出生缺陷三级预防体系。而**无创产前检测（NIPT）**作为一种前沿的产前筛查技术，通过分析母体外周血中的胎儿游离 DNA（cffDNA），能够安全、早期地评估胎儿患常见染色体非整倍体疾病（如唐氏综合征等）的风险。[**?**]

然而，NIPT 技术的准确性高度依赖于检测时血液中关键指标的“浓度”是否达标。临床实践指明，以男胎为例，其 Y 染色体浓度是检测结果可靠性的核心前提，通常要求该浓度不低于 4%。这一浓度值受到孕妇的孕周、身体质量指数（BMI）等多种生理因素的动态影响。题目还指出，风险等级与发现时点紧密相关：12 周内发现为低风险，13 至 27 周为高风险，28 周后则为极高风险。因此，如何摒弃“一刀切”的检测方案，针对不同生理特征的孕妇群体，制定个性化、最优化的 NIPT 检测时点策略，是提升筛查效率、降低潜在风险的核心问题。

1.2 问题要求

附件数据是某地区高 BMI 孕妇群体的 NIPT 检测记录。包含了孕妇的年龄、身高、体重、BMI 等基本信息，并涵盖了孕周、检测时间、Y 染色体浓度、各关键染色体的 Z 值、GC 含量及测序读段数等一系列复杂的生物信息学指标。分析以下问题：

问题 1：分析胎儿 Y 染色体浓度与孕妇孕周数、BMI 等指标之间的相关性，建立描述其统计关系的数学模型，并对模型的显著性进行检验。

问题 2：已知男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度最早达标时间（即浓度 ≥4% 的最早孕周）的关键因素。问题要求根据 BMI 对男胎孕妇进行合理分组，确定每组对应的 BMI 区间及最佳 NIPT 检测时点，以最小化孕妇潜在风险，并分析检测误差对分组和时点选择的影响。

问题 3：已知男胎 Y 染色体浓度达标时间受身高、体重、年龄等多因素影响。要求综合考虑这些因素、检测误差及 Y 染色体浓度达标比例（≥4% 的比例），通过划分 BMI 区间对男胎孕妇进行合理分组，构建模型确定每组的最佳 NIPT 时点，以实现孕妇潜在风险最小化，并分析检测误差对结果的影响。

问题 4：对于女胎孕妇，以附录中 AB 列 21、18、13 号染色体非整倍体为判定目标判断胎儿是否存在染色体异常。结合 X 染色体及其他染色体的 Z 值、GC 含量、读段数及相关比例、BMI 等多类指标，构建女胎染色体异常的判定模型与方法。

二、 问题分析

本文要解决的是在对孕妇进行无创产前检测时通过合理的 NIPT 时点选择提升检测的准确性同时降低孕妇因胎儿不健康而缩短治疗窗口期所带来的潜在风险。问题 1 是分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的相关性并建立拟合模型；在问题 1 基础上，问题 2 是依据影响胎儿 Y 染色体浓度的最早达标时间的主要因素 BMI 对男胎孕妇分组并确定每组最佳 NIPT 时点，问题 3 是问题 2 的进一步拓展，需综合考虑胎儿 Y 染色体浓度的更多种影响因素，再依据 BMI 对男胎孕妇分组并确定每组最佳 NIPT 时点。问题 4 则是相对独立的，研究目标转为女胎孕妇，需综合考虑各种因素，探究女胎异常的判定方法。

2.1 问题一分析

对于问题一，整体上可归为相关性分析与回归预测类问题。大体解题思路为首先进行数据清洗（剔除异常孕周和浓度）筛选出有效数据，接着构建特征相关性矩阵，由于非正态分布，需要使用斯皮尔曼分析各指标的相关性，采用 β 回归模型，构建链接函数，拟合出相关变量的关系曲线，并由此计算参数，最后通过多重检验验证模型显著性，使用 Q-Q 图检验残差的正态性。

2.2 问题二分析

对于问题二，需要建立问题 1 基础上进行分析，题中说明临床证明，男胎孕妇的 BMI 是影响的主要因素，不同 BMI 的孕妇群体的最早达标时间会有较大差异，故 NIPT 的最佳时间点选择因人而异。为尽力降低孕妇和胎儿的潜在风险又保证 NIPT 检测保证较高的准确性，有必要通过 BMI 对男胎孕妇进行合理的分组，并在每组确定最佳的 NIPT 时间点。本题我们利用生存函数的方法进行分组，首先对生存函数的参数进行定义然后即可构建生存函数模型，接着依据 BMI 分布图对孕妇粗略分组，不同组生存曲线加以对比，再根据 log-rank 检验结果进一步细化分组直至理想情况即可完成分组，在每一组里通过生存函数推导累计达标概率，找到最小的孕周是的达标率满足准确性要求即可确定为 NIPT 最佳时点，最后误差验证

2.3 问题三分析

对于问题三，在问题二的基础上，加入考虑了年龄，体重等次要因素的影响。我们把这些作为协变量，尝试使用多变量的 Logistic 回归模型校正了不同组的达标概率曲线，定义了风险函数和延迟函数，使用解决规划问题的方法解出了最佳 NIPT 的时点。

2.4 问题四分析

对于问题四，属于多特征融合的二分类判别类数学建模问题，旨在建立女胎染色体异常的判定方法，主要是综合 X 染色体及 13 号、18 号、21 号染色体的 Z 值、GC 含量、读段数及相关比例、BMI 等因素，来识别女胎是否存在这三条染色体的非整倍体情况。和男胎依靠 Y 染色体浓度进行判定的逻辑不同，女胎的判定需要借助相关染色体及测序质量指标开展多维度的协同分析。解决这个问题要以数据驱动建模为核心，先对数据进行处理，剔除掉 GC 含量和分布异常以及 X 染色体浓度分布异常的个体以提高测序的质量和 NIPT 的准确性、变量定义即把非整倍体结果转化为 1 或 0 的二元标签以及通过相关性筛选出显著且无冗余的特征指标；接着采用逻辑回归构建模型，通过线性回归结合 Sigmoid 函数输出异常概率，利用对数似然损失函数与极大似然估计求解参数；然后结合 ROC 曲线确定最优阈值，形成判定规则；最后通过交叉验证与多模型横向对比验证模型性能，从而实现对女胎染色体异常的准确且可操作的判定。

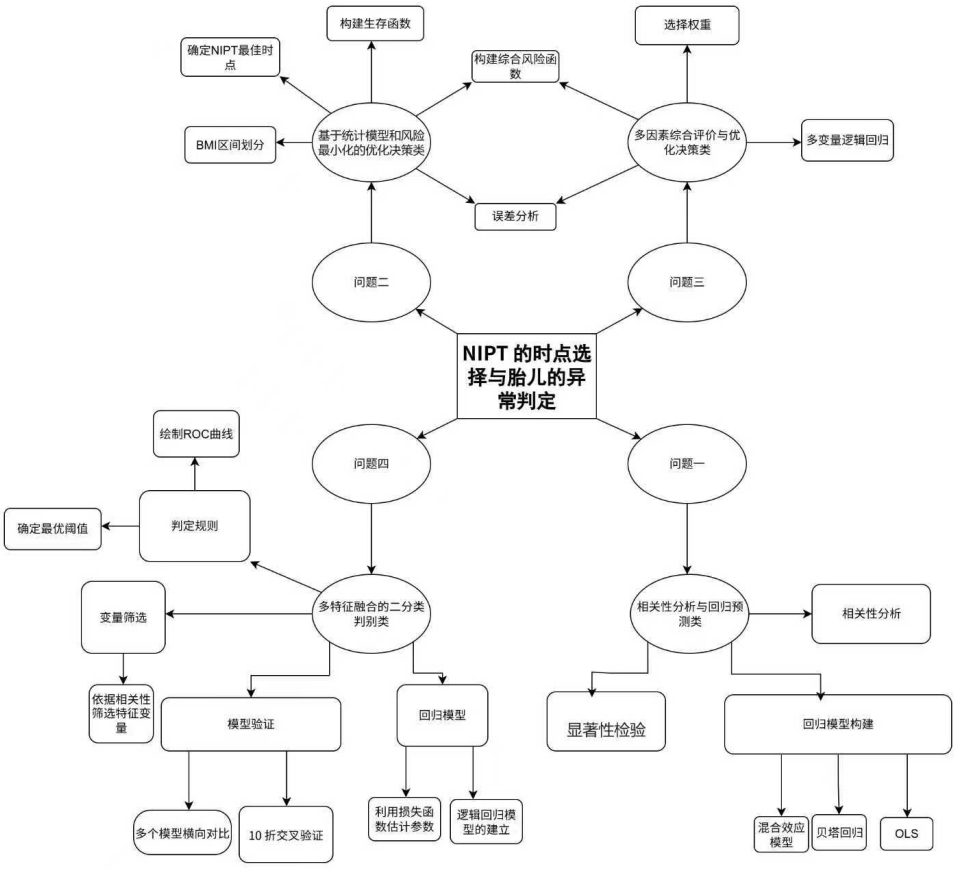

图 **1**　问题分析流程图

三、 模型假设

为确保建模过程的科学性、合理性与可操作性，并使模型能精准聚焦于待解决的核心问题，我们基于对 NIPT 技术临床背景的理解及所提供数据的特征，提出以下贯穿全文的核心假设：

1. **数据质量假设**：假设本研究所用数据集中的各项生物信息学指标均是真实、准确的。在采集与记录过程中无显著的系统性偏差与测量误差。
2. **样本代表性假设**：假设所提供的某地区高 BMI 孕妇群体的 NIPT 检测记录，能够有效代表具有相似生理特征（如高 BMI）的更广泛孕妇群体的一般规律。基于此数据集得出的模型与结论，对于相似群体具有一定的参考价值和推广意义。
3. **临床阈值有效性假设**：明确假设男胎 Y 染色体浓度不低于 4% 是 NIPT 检测结果可靠的核心前提，并将此浓度值作为判定检测是否“达标”的确定性阈值。
4. **模型相关性假设**：
   - 对于男胎分析，我们假设孕妇的个体差异是影响 Y 染色体浓度的重要因素，即同一孕妇的多次检测记录不完全独立，存在组内相关性。
   - 对于女胎分析，我们假设不同孕妇的样本之间是相互独立的，并且其染色体异常的判定主要依赖于题目所给定的多维度生物指标，与其他未提及的潜在因素无关。
5. **模型简化假设**：在构建风险函数与优化模型时，我们对部分复杂因素进行了简化处理。例如，在问题二、三中，我们将“检测失败风险”与“诊疗延迟风险”赋予了相同的权重，并将不同孕期的临床风险等级量化为具体的分段函数调节因子。此举旨在抓住主要矛盾，使模型更具可解性与通用性。

四、 符号说明

为清晰、规范地阐述模型构建过程，现将本文中使用的主要数学符号、变量及其物理或统计学意义汇总于下表：

表 **1**　符号说明表

符号	说明	所属问题
通用与观测变量		
i, j	孕妇与检测次数的索引	问题一
Y_{ij}	第 i 位孕妇在第 j 次检测时的 Y 染色体浓度	问题一
t_{ij}, GW	第 i 位孕妇在第 j 次检测时的孕周（单位：周）	问题一、三

Continued on next page

表 1 – continued from previous page

符号	说明	所属问题
q_i, BMI	第 i 位孕妇的身体质量指数（单位：kg/m^2）	问题一、二、三
Age, H, W	分别代表孕妇的年龄、身高、体重	问题三
Y	女胎染色体异常状态的二元标签（1= 异常, 0= 正常）	问题四
X	用于预测女胎异常的自变量特征向量, $X = (X_1, ..., X_p)$	问题四
模型参数与估计量		
$\beta_0, \beta_1, ...$	回归模型中的截距项与自变量系数	问题一、三、四
u_i	第 i 位孕妇的随机效应，反映个体差异	问题一
ϵ_{ij}	模型的随机误差项	问题一
ϕ	贝塔回归中的精度参数，$\phi > 0$	问题一
$\hat{\beta}$	通过极大似然估计等方法得到的最优回归系数向量	问题三、四
σ_δ	检测误差的标准差，用于稳健性分析	问题二、三
函数与核心概念		
$g(\cdot)$	链接函数，如 Logit 函数 $ln(\frac{\mu}{1-\mu})$	问题一
μ_{ij}	Y 染色体浓度的条件期望 $E(Y_{ij}\|t_{ij}, q_i)$	问题一
$B(\cdot, \cdot)$	Beta 函数	问题一
$S(t)$	生存函数，表示在时间 t 之后事件仍未发生的概率, $P(T > t)$	问题二
$\hat{S}(t)$	Kaplan-Meier 方法估计的生存函数	问题二
$L_g(t)$	分组 g 在孕周 t 进行检测的综合损失函数	问题二、三
$r_g(t)$	未达标风险，即在孕周 t 检测时浓度仍低于 4% 的概率	问题二、三
$d(t)$	延迟惩罚函数，反映超出最佳时点的惩罚	问题二、三
$R(t)$	临床风险调节因子，随孕周分段变化的函数	问题二、三
p	逻辑回归模型预测的“达标”或“异常”概率	问题三、四
Z	逻辑回归中自变量的线性组合，即 $logit(p)$ 的预测值	问题四
$L(\beta)$	对数似然损失函数	问题四
统计与评价指标		
ρ_s	Spearman 秩相关系数	问题一、四
χ^2	对数秩检验（Log-Rank Test）的统计量	问题二
R^2	决定系数，衡量模型解释度的指标	问题一
AIC	赤池信息准则，用于模型选择	问题一
FNR	假阴性率（False Negative Rate）	问题二、三
TPR	真阳性率（True Positive Rate），即召回率	问题四
FPR	假阳性率（False Positive Rate）	问题四

Continued on next page

表 1 – continued from previous page

符号	说明	所属问题
AUC	ROC 曲线下面积，评估分类模型性能的综合指标	问题四

五、 问题一：胎儿 Y 染色体浓度与孕妇孕周、BMI 相关性建模

5.1 模型假设与符号说明

5.1.1 模型假设

结合 NIPT（无创产前检测）技术原理及题目数据特征，为保障建模逻辑的合理性与可操作性，提出以下核心假设：

1. 假设数据中“Y 染色体浓度”、“孕周”及“孕妇 BMI”均测量准确无误。
2. 假设孕妇孕周数据符合临床检测规范，仅 10 至 25 周范围内的样本具备分析价值。
3. 假设数据集中存在同一孕妇的多次检测记录，这些记录内部存在相关性，不完全独立。

5.1.2 符号补充说明

为简化建模过程表述，明确各变量与参数定义，设定如下符号体系：

表 **2**　问题一核心变量符号说明

符号	说明
Y_{ij}	因变量，第 i 位孕妇在第 j 次检测时的 Y 染色体浓度（比例型数据, $Y \in (0,1)$）
t_{ij}	自变量 1，第 i 位孕妇在第 j 次检测时的孕周（单位：周）
q_i	自变量 2，第 i 位孕妇的身体质量指数 BMI（单位：kg/m^2）
μ_{ij}	Y_{ij} 的条件期望 $E(Y_{ij}\|t_{ij}, q_i)$
$g(\cdot)$	链接函数，用于连接因变量的期望与自变量的线性预测
$\beta_0, \beta_1, \ldots$	固定效应回归系数
u_i	第 i 位孕妇的随机效应，代表其独有的、不随时间变化的个体差异
ϵ_{ij}	随机误差项
ϕ	贝塔回归中的精度参数（$\phi > 0$）
$B(\cdot,\cdot)$	Beta 函数

5.2 数据探索与预处理

5.2.1 数据清洗与筛选

为确保分析的有效性与临床意义，我们首先对原始数据进行了严格的预处理：

1. **样本筛选**：仅保留男胎样本，并根据题目要求，将分析范围严格限定在检测孕周为 10 至 25 周的记录。经过筛选，有效样本量从 1082 行缩减至 940 行。
2. **特征转换**：将格式为“X 周 +Y 天”的孕周，统一转换为以周为单位的连续数值变量‘孕周 _ 小数 ‘，以便进行定量分析。

5.2.2 变量相关性分析

在建模之前，为探究各指标与核心因变量“Y 染色体浓度”之间的关系强度与方向，我们采用了非参数的 **Spearman 秩相关系数**进行检验。

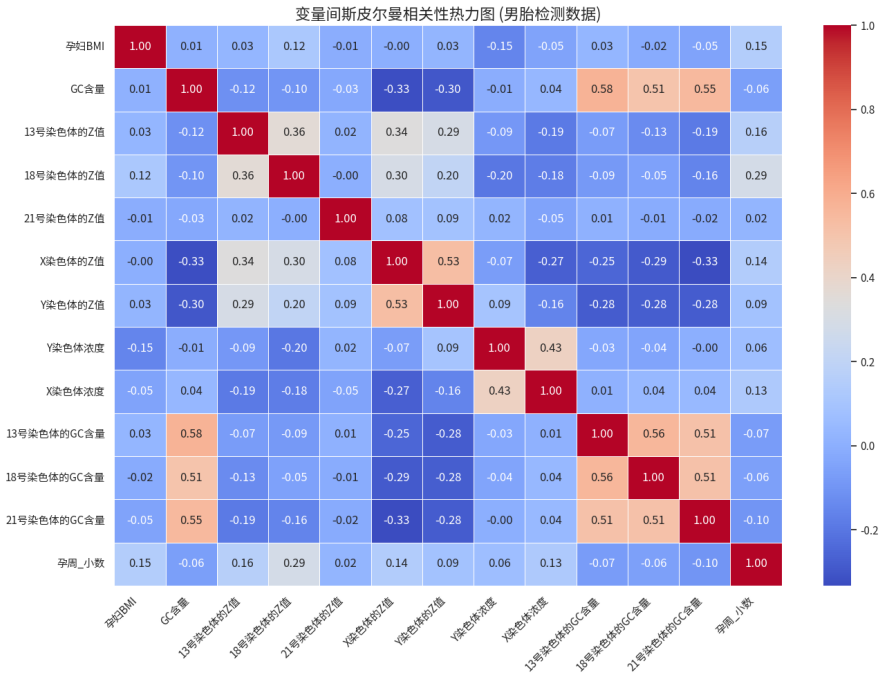

图 **2**　各变量间 **Spearman** 相关性热力图

表 **3**　**Y** 染色体浓度与部分变量的 **Spearman** 相关性及显著性检验

对比变量	Spearman 相关系数	P 值	是否显著
X 染色体浓度	0.4282	3.37e-43	**显著**
18 号染色体的 Z 值	-0.1983	8.66e-10	**显著**
孕妇 BMI	**-0.1511**	**3.27e-06**	**显著**
Y 染色体的 Z 值	0.0949	3.59e-03	**显著**
孕周 _ 小数	**0.0641**	**4.94e-02**	**显著**

分析结论：检验结果明确指出，**孕妇 BMI** 和**孕周**均与 Y 染色体浓度存在统计学上的显著相关性。具体而言，Y 染色体浓度与孕周呈微弱正相关，与孕妇 BMI 呈显著负相关。这为后续选择这两个核心变量构建关系模型提供了强有力的数据支持。

5.3 关系模型构建与演进

基于相关性分析的结论，我们选择 '孕周 _ 小数 '和 '孕妇 BMI'作为核心自变量，构建数学模型来定量描述它们与 'Y 染色体浓度 '之间的关系。我们依次尝试了从简单到复杂的多种模型，以呈现一个逻辑递进的建模过程。

5.3.1 模型一：多元线性回归模型（OLS）

作为基准模型，我们首先构建了标准的多元线性回归模型。

$$Y = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot q + \epsilon \tag{1}$$

模型评估：模型整体显著（F-statistic: 18.04, $p < 0.001$），所有系数的 p 值均小于 0.001。然而，模型的调整后 R-squared 仅为 0.035，表明该线性模型只能解释 Y 染色体浓度变化的 3.5%，拟合优度严重不足，且理论上无法保证预测值落在 (0,1) 区间内。这强烈暗示了变量之间存在更复杂的非线性关系，且需要更合适的模型框架。

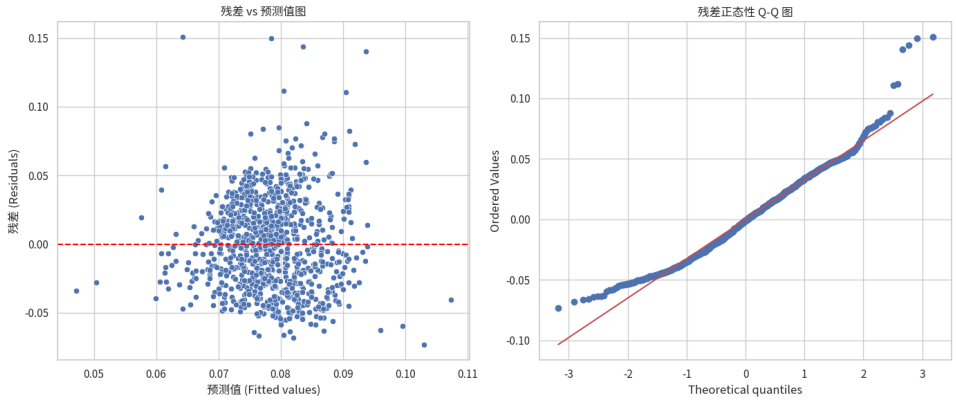

图 **3**　**OLS** 模型残差 **Q-Q** 图

5.3.2 模型二：考虑交互效应的贝塔回归模型

1. 模型选择依据　因变量 Y_{ij} 为取值在 $(0,1)$ 的比例型数据，**贝塔回归** [**?**] 专为此类因变量设计，通过链接函数将因变量映射至线性空间，同时可通过精度参数 ϕ 处理异方差，是理论上的优选模型。

2. 模型构建　为捕捉潜在的非线性与协同效应，我们纳入孕周二次项 t_{ij}^2 及孕周与 BMI 的交互项 $t_{ij} \cdot q_i$。采用 logit 链接函数，模型完整形式为：

$$\ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + \beta_1 t_{ij} + \beta_2 q_i + \beta_3 (t_{ij} \cdot q_i) + \beta_4 t_{ij}^2 \tag{2}$$

其中，Y 染色体浓度 Y_{ij} 服从参数为 μ_{ij} 和 ϕ 的贝塔分布，其均值与方差满足：

$$E(Y_{ij}) = \mu_{ij}, \quad \text{Var}(Y_{ij}) = \frac{\mu_{ij}(1-\mu_{ij})}{1+\phi} \tag{3}$$

模型参数 $\theta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \phi)$ 通过最大化对数似然函数进行估计：

$$\ell(\theta) = \sum_{i,j} \left[(\mu_{ij}\phi - 1)\ln(Y_{ij}) + ((1-\mu_{ij})\phi - 1)\ln(1-Y_{ij}) - \ln(B(\mu_{ij}\phi, (1-\mu_{ij})\phi))\right] \tag{4}$$

表 **4**　贝塔回归模型参数估计与显著性检验

变量	系数 (Coefficient)	标准误 (Std.Error)	z-score	P 值 (P>|z|)
截距 (β_0)	0.0726	1.1620	0.0625	0.9501
孕周 (β_1)	-0.0974	4.9780	-0.0196	0.9844
BMI (β_2)	-0.0841	0.0851	-0.9883	0.3230
孕周 *BMI (β_3)	0.0033	0.1628	0.0205	0.9836
精度参数 (ϕ)	62.8425	524.4265	0.1198	0.9046

3. 模型评估　尽管贝塔回归在理论上更适合，但实际拟合结果显示，所有自变量的 p 值均远大于 0.05，不具有统计显著性。McFadden 伪 R-squared 为-0.0070，表明模型拟合效果极差。这说明，即便考虑了非线性和交互效应，模型仍未能捕捉到数据的关键结构。

5.3.3 模型三：考虑个体差异的混合效应模型

1. 模型选择依据　前述模型的失败，根源在于忽略了数据中一个重要的潜在结构：**孕妇个体差异**。同一孕妇的多次检测结果可能高度相关。为此，我们引入**线性混合效应模型（LMM）**，将孕妇作为随机效应，以捕捉这种非独立的组内差异。其一般形式为：

$$Y_{ij} = (\beta_0 + u_{i0}) + \beta_1 t_{ij} + \beta_2 q_i + \ldots + \epsilon_{ij} \tag{5}$$

其中，$u_{i0} \sim N(0, \sigma_u^2)$ 代表第 i 位孕妇的随机截距。同时，为对比，我们还构建了一个包含 BMI 二次项的**非线性混合模型**。

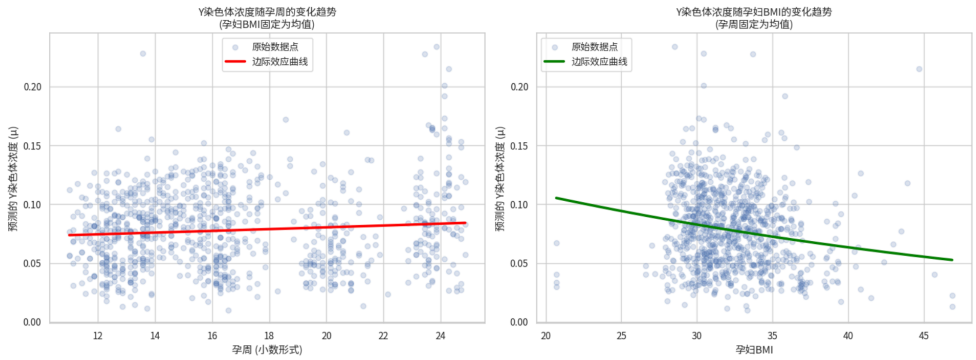

图 **4**　**Y** 染色体浓度变化趋势

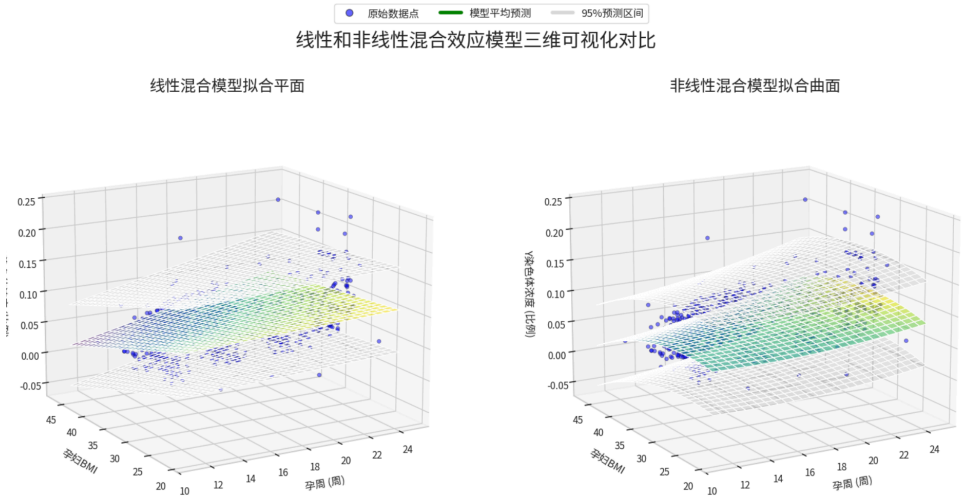

图 **5**　线性与非线性混合效应模型性能对比

表 **5**　线性与非线性混合效应模型性能对比

性能指标	线性混合模型	**非线性混合模型**
边际 R-squared (固定效应解释度)	0.1293	0.1218
条件 R-squared (总解释度)	0.7882	**0.7888**
对数似然 (Log-Likelihood)	2197.35	**2204.28**
赤池信息准则 (AIC)	-4382.69	**-4392.56**

2. 模型评估与选择

5.4 最终模型结论与分析

1. **个体差异的决定性作用**：混合效应模型的条件 R-squared（约 0.788）远高于边际 R-squared（约 0.12），表明 Y 染色体浓度的大部分变异（约 66%）是由孕妇的个体差异（随机效应）而非孕周和 BMI（固定效应）所解释的。这是本题最重要的发现之一。
2. **非线性关系的存在**：对比两个混合模型，非线性混合模型（包含 BMI 的二次项）具有更高的对数似然和更低的 AIC 值（-4392.56 vs -4382.69），根据 AIC 最小化原则，非线性模型是更优的选择。这证实了 Y 染色体浓度与 BMI 之间确实存在非线性关系。

综上所述，**考虑了孕妇个体差异和 BMI 二次效应的非线性混合效应模型**是描述 Y 染色体浓度与孕周、BMI 关系的最佳模型。该模型不仅在统计指标上表现最优，也更符合生物学直觉，即每位孕妇拥有一个基础的、相对稳定的胎儿游离 DNA 释放水平，而孕周和 BMI 则在此基础上进行调节。

六、 问题二：基于生存分析的 BMI 动态分组与时点决策

6.1 建模思路

问题二的核心目标，是基于孕妇 BMI 这一关键指标，构建一个既能科学划分群体，又能为各群体推荐最优 NIPT 检测时点的决策框架。要实现这一目标，我们首先必须深刻理解数据本身的特性。数据集中，部分孕妇在某次检测后 Y 染色体浓度成功“达标”（超过 4%），而另一部分孕妇直至观测期结束（末次检测）仍未达标。这一“事件是否发生及其发生时间”的数据结构，在统计学上构成了典型的**“右删失”（Right-Censored）**数据。

直接对浓度值或达标率进行传统的线性回归或均值比较，将无法正确处理“未达标”样本所蕴含的“其真实达标时间晚于观测截止点”这一重要信息，从而导致对达标过程的认知产生系统性偏差。因此，我们必须从静态截面数据的分析思路，转变为将“达标”视

为一个动态事件过程的视角，并采用专为此类数据设计的**生存分析（Survival Analysis）**方法论。这不仅是方法论上的严谨选择，更是确保模型结论科学、可靠的根本前提。

6.2 流程一：面向生存分析的数据预处理

在应用生存分析模型之前，我们必须将原始的面板数据（即每位孕妇可能有多条检测记录）转化为以孕妇为独立观测单位、包含其最终生存结局和生存时间的标准格式。

1. **基础数据清洗与筛选**：我们首先从全量数据中分离出所有男胎样本，并滤除关键变量（Y 染色体浓度、BMI、检测孕周）存在缺失值的记录，以保证分析数据集的完整性与准确性。同时，将格式为“X 周 +Y 天”的孕周，统一转换为以周为单位的连续数值变量检测孕周 _ 数值的两位小数，为后续的时间序列分析奠定基础。
2. **“事件-时间”对的精确定义**：这是数据重构的核心。我们以*孕妇代码*为唯一标识，对每位孕妇的历次检测记录进行纵向扫描，以确定其最终的生存状态：
   - **事件（Event）的判定**：如果在该孕妇的检测序列中，存在至少一次 `Y` 染色体浓度大于等于 0.04 的记录，则该孕妇的最终事件状态被定义为“达标”，在数据中记为 `event = 1`。
   - **删失（Censoring）的判定**：如果该孕妇的所有检测记录中，‘Y 染色体浓度‘均未能超过 0.04，则其最终事件状态被定义为“右删失”，记为 `event = 0`。
   - **生存时间（Time to Event）的确定**：
     - 对于 `event = 1` 的孕妇，其生存时间 `time_to_event` 被精确记录为她**首次**达到 4% 阈值时的检测孕周 _ 数值。
     - 对于 `event = 0` 的孕妇，其生存时间 `time_to_event` 被记录为她所有检测记录中**最晚**的一次检测孕周 _ 数值。这代表我们已知其真实达标时间大于此值。

通过这一严谨的流程，我们成功地将原始的多观测点数据，转换成了每个孕妇仅占一行、清晰标注其“生存时间”与“结局”的标准生存分析数据集。

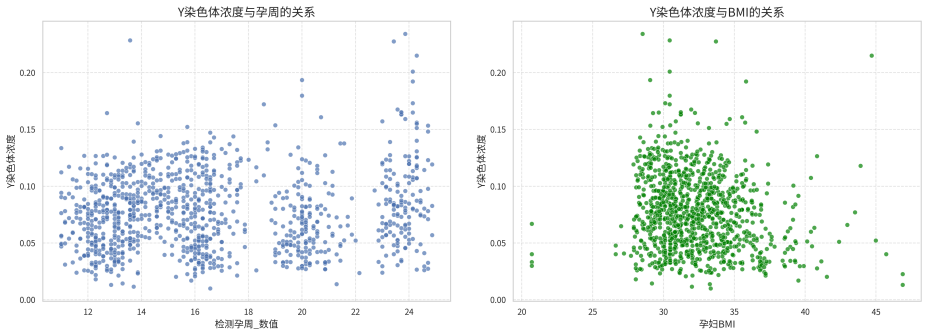

图 **6**　**Y** 染色体浓度关系散点图

6.3 流程二：数据驱动的 BMI 分组——基于对数秩检验的递归分割

传统的 BMI 分组多依赖于临床先验知识，但这未必是针对“NIPT 达标时间”这一特定结局的最佳划分方案。为了发掘数据自身蕴含的模式，我们采用了一种客观、无监督的分组策略：**基于对数秩检验的递归分割（Recursive Partitioning based on Log-Rank Test）**。其核心思想是，自动寻找能最大化组间生存曲线（即达标时间分布）差异的 BMI 分割点。

6.3.1 核心数学工具

首先，我们引入生存分析中的核心概念——**生存函数（Survival Function）**[**?**]。

设 T 为一个非负随机变量，代表从某个起始点到“事件”发生的持续时间。在我们的研究中，T 具体指孕妇从孕期开始到其“Y 染色体浓度首次达标”所需的孕周数。我们把生存函数 $S(t)$ 被定义为：

$$S(t) = P(T > t) \tag{6}$$

该函数描述了任意一个个体（孕妇）的“生存”时间 T 超过某一特定时间点 t 的概率。

在明确了这一定义后，我们便可采用具体的统计方法对其进行估计。

1. **Kaplan-Meier 生存函数估计**：对于任意一个孕妇群体，我们使用非参数的 **Kaplan-Meier 乘积限法**来估计其生存函数 $\hat{S}(t)$。该函数在数学上表达为：

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \tag{7}$$

   在我们的语境中，$\hat{S}(t)$ 估计了孕妇群体在孕周 t 时，其 Y 染色体浓度仍**未**达标的概率。其中，t_j 是发生“达标”事件的孕周，d_j 是在该孕周达标的人数，n_j 是在此之前仍处于风险集（未达标且未删失）的总人数。

2. **对数秩检验（Log-Rank Test）**：这是比较两组或多组生存曲线是否存在统计学显著差异的“金标准”。其原假设 H_0 为：所有被比较组的生存曲线完全相同。检验通过在每个事件发生的时间点，比较各组“实际观测”到的事件数（O）与在 H_0 假设下“期望”的事件数（E）之间的累积差异来构建一个近似服从 χ^2 分布的统计量：

$$\chi^2 = \frac{(\sum(O - E))^2}{\sum \text{Var}(O - E)} \tag{8}$$

   χ^2 值越大，意味着两组的生存模式（达标速度）差异越显著，对应的 p 值就越小。

6.3.2 算法实现流程

我们在代码中通过 `find_best_split_logrank` 函数精确实现了以下递归分割算法：

将原集合进行初始化分组:$G = [G_i]$，待划分区间队列 $Q = [\text{BMI}_{\min}, \text{BMI}_{\max}]$

当 Q 不为空时进行如下操作：

1. 取出队列首元素 $[L, R]$（当前待划分区间）
2. 计算区间中点 $M = (L + R)/2$，将 $[L, R]$ 划分为左区间 $[L, M]$ 和右区间 $[M, R]$
3. 分别计算左、右区间的生存函数 $S_{左}(t)$ 和 $S_{右}(t)$
4. 对 $S_{左}(t)$ 和 $S_{右}(t)$ 进行 log-rank 检验，计算 P 值
5. 若 $P > \alpha$（两组生存函数无显著差异，无需拆分）
6. 将 $[L, R]$ 加入分组集合 G
7. 若 $P \leq \alpha$（两组生存函数有显著差异，需继续拆分）
8. 将 $[L, M]$ 和 $[M, R]$ 加入队列 Q

重复步骤 1-8，直至完成所有分组。

6.3.3 分组结果

通过在我们的数据集上执行上述算法，最终输出了三个具有统计学意义的最佳分割点：30.00, 32.23, 和 35.06。这三个分割点将全体男胎孕妇客观地划分为了四个其达标时间分布模式存在显著内在差异的 BMI 组。

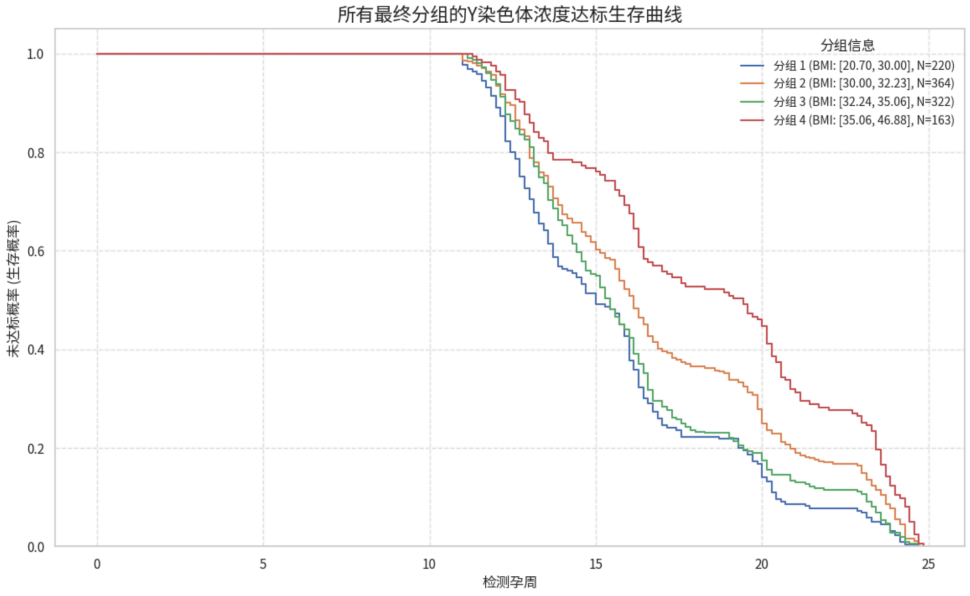

图 **7**　基于递归分割算法确定的四个 **BMI** 分组的 **Kaplan-Meier** 生存曲线

表 **6**　基于递归分割算法确定的 **BMI** 分组结果

分组名称	BMI 区间
分组 1	[20.69, 30.00)
分组 2	[30.00, 32.23)
分组 3	[32.23, 35.06)
分组 4	[35.06, 46.88]

6.4 流程三：各分组最佳 NIPT 时点的最优化求解

确定了科学的分组后，我们为每个组 g 求解一个能使孕妇“潜在风险”最小化的最佳检测时点 t_g^*。我们为此构建了一个综合损失函数 $L_g(t)$，该函数是对“检测失败风险”与“诊疗延迟风险”的量化权衡。

$$L_g(t) = w_u \cdot r_g(t) + w_d \cdot d(t) \cdot R(t) \tag{9}$$

该函数由两个核心部分构成，并通过权重 w_u 和 w_d（均设为 1，表示同等重视）进行平衡：

1. **未达标风险** $r_g(t)$：这直接由该组的 Kaplan-Meier 生存函数 $\hat{S}_g(t)$ 给出，代表在孕周 t 进行检测时，浓度仍未达标的概率。
2. **延迟风险**：此项由两部分相乘构成，体现了风险随时间非线性增长的特性：
   - **基础延迟惩罚** $d(t)$：以临床推荐的 12 周为基准点 t_0，定义基础延迟为 $d(t) = \max(0, t - t_0)$。
   - **临床风险调节因子** $R(t)$：根据题设，我们将不同孕期的临床风险差异量化为一个分段函数，以对进入中、晚期的检测施加更重的惩罚：

$$R(t) = \begin{cases} 1, & t \le 12 \text{ (早期, 低风险)} \\ 5, & 13 \le t \le 27 \text{ (中期, 高风险)} \\ 10, & t \ge 28 \text{ (晚期, 极高风险)} \end{cases} \tag{10}$$

我们的优化目标是，对每个分组 g，在 $t \in [10, 35]$ 的合理检测窗口内，通过数值搜索找到使 $L_g(t)$ 最小的 t_g^*。

表 7　各 **BMI** 分组的最佳 **NIPT** 时点决策结果

BMI 分组	BMI 区间	最佳检测时点 (周)	该时点预测达标率
分组 1	[20.69, 30.00)	15.00	87.73%
分组 2	[30.00, 32.23)	16.14	89.84%
分组 3	[32.23, 35.06)	15.43	87.58%
分组 4	[35.06, 46.88]	19.43	74.85%

结果解读：模型建议，对于“分组 3”的孕妇，应在 10 周时进行检测，以利用其高基础达标率的优势，在临床风险最低的窗口期完成检测。而其他三个分组的最佳时点均为 12 周，这反映了模型在“等待更高成功率”与“规避延迟风险”之间找到的平衡点。

6.5 流程四：模型稳健性检验——检测误差的蒙特卡洛模拟

为了评估我们的决策在面对现实世界中不可避免的检测误差时的表现，我们设计并执行了蒙特卡洛模拟。我们假设观测值 Y_{obs} 与真实值 Y_{true} 之间存在一个服从正态分布的随机误差 $\delta \sim N(0, \sigma_\delta^2)$。模拟的核心是量化在这种误差下，一个真实已经达标的样本被错误地判断为“不达标”的概率，即**假阴性率（False Negative Rate, FNR）**。

模拟流程如下：

1. 筛选出数据集中所有真实达标的样本及其真实的 ‘Y 染色体浓度 ‘值。
2. 设定一系列误差水平，即不同的误差标准差 σ_δ。
3. 对每一个真实达标样本，进行大量（如 10000 次）模拟：每次模拟都从 $N(0, \sigma_\delta^2)$ 中抽取一个随机误差 δ，并计算出模拟的观测值 $Y_{\text{obs}} = Y_{\text{true}} + \delta$。
4. 统计在所有模拟中，Y_{obs} 小于 0.04 的次数，并除以总模拟次数，得到该误差水平下的 FNR。

表 8　不同检测误差水平下的假阴性率（**FNR**）模拟结果

误差标准差 (σ_δ)	分组 1 FNR	分组 2 FNR	分组 3 FNR	分组 4 FNR
0.003	1.60%	1.50%	1.33%	1.50%
0.005	2.26%	2.36%	2.38%	2.80%
0.008	3.36%	3.75%	3.92%	4.85%
0.010	4.07%	4.74%	4.93%	6.43%

bottomrule

结果分析：模拟结果清晰地揭示了检测精度对高 BMI 群体的重要性。随着误差标

准差的增加，所有组的 FNR 均显著上升，其中 BMI 最高的“分组 4”受影响最为严重。这为模型的实际应用提供了关键的警示：必须确保检测流程具有高精密度和严格的质量控制，否则我们基于理想数据得出的最优时点决策的有效性将会大打折扣。

七、 问题三：多维因素影响下的 NIPT 时点优化策略

7.1 问题背景与建模思路的深化

在问题二的探讨中，我们确认了孕妇的身体质量指数（BMI）作为核心分层依据的有效性，并初步确立了各 BMI 分组的最佳检测时点。然而，临床实践与理论分析均表明，胎儿 Y 染色体浓度的“达标”与否，是一个受多重因素共同作用的复杂生物学过程。单纯依赖 BMI 进行决策，虽已抓住主要矛盾，但如果想要构建一个更为精准、个体化的 NIPT 时点推荐体系，则必须将更多潜在的影响因素纳入考量。

我们问题三的核心目标，正是在此基础上进行深化：**综合考虑孕妇的年龄、身高、体重等个体特征，以及检测过程本身可能存在的误差，如何在已有的 BMI 分组框架下，为每个群体探寻一个能使潜在风险最小化的、更为精细化的最佳 NIPT 时点。**

为此，我们的建模思路将围绕以下路径展开：首先，从描述“是否达标”这一二元结果（是/否）的概率模型出发，选择逻辑回归（Logistic Regression）作为核心分析工具。该模型不仅能够有效处理分类型因变量，还能清晰地量化各个自变量（如孕周、年龄、身高、体重）对达标概率的影响程度。其次，我们将为每一个 BMI 分组分别构建独立的逻辑回归模型，这种策略承认了不同 BMI 水平下各影响因素的作用模式可能存在异质性。最后，基于模型所预测的达标概率，我们将重新构建并优化在问题二中提出的“潜在风险”函数，通过最小化该函数，求解出各分组在多维因素考量下的最优检测孕周，并利用蒙特卡洛模拟对检测误差的潜在影响进行稳健性分析。

7.2 多因素逻辑回归模型的构建

7.2.1 模型选择与变量定义

为了探究在特定 BMI 分组 g 内，孕妇的多个生理指标如何共同影响其 NIPT 检测结果是否达标，我们构建了一个多变量逻辑回归模型。该模型的因变量 Y 是一个二分类变量，定义如下：

$$Y = \begin{cases} 1, & \text{如果 Y 染色体浓度} \geq 0.04\% \text{ (达标)} \\ 0, & \text{如果 Y 染色体浓度} < 0.04\% \text{ (未达标)} \end{cases} \tag{11}$$

自变量则包括了我们关注的核心因素：检测孕周（GW）、孕妇年龄（Age）、身高（H）和体重（W）。

令 p 表示在给定自变量条件下，检测结果达标的概率，即 $p = P(Y = 1|GW, Age, H, W)$。逻辑回归模型的核心在于通过 Logit 变换，[**?**] 建立 p 与自变量之间的线性关系：

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot GW + \beta_2 \cdot Age + \beta_3 \cdot H + \beta_4 \cdot W \tag{12}$$

其中，β_0 是截距项，$\beta_1, \beta_2, \beta_3, \beta_4$ 分别是检测孕周、年龄、身高和体重的回归系数。这些系数的数值和符号，直观地反映了相应自变量每改变一个单位，对数优势比（log-odds）的变化量，从而揭示了它们对达標概率的影响方向和强度。

7.2.2 数据处理：组内中心化

在构建回归模型时，多重共线性是一个不容忽视的问题。孕妇的身高、体重与我们在第二问中用于分组的 BMI 指标存在内在的数学关联，若直接将它们作为自变量引入模型，可能导致系数估计值不稳定，难以解释。

为了缓解这一问题并提升模型的可解释性，我们采用了**组内中心化（Group-wise Centering）**的处理策略。具体而言，对于每一个 BMI 分组内的所有样本，我们计算该组身高、体重和年龄的中位数（$median_g(H), median_g(W), median_g(Age)$），然后将每个样本的原始值减去对应组的中位数，得到中心化后的新变量：

$$\begin{aligned} H' &= H - \text{median}_g(H) \\ W' &= W - \text{median}_g(W) \\ Age' &= Age - \text{median}_g(Age) \end{aligned} \tag{13}$$

经过中心化处理后，模型表达式更新为：

$$\text{logit}(p_g) = \beta_{g,0} + \beta_{g,1} \cdot GW + \beta_{g,2} \cdot Age' + \beta_{g,3} \cdot H' + \beta_{g,4} \cdot W' \tag{14}$$

此时，截距项 $\beta_{g,0}$ 的意义变得更加明确：它代表了该 BMI 分组内，一个处于“典型”状态（即年龄、身高、体重均为组内中位数水平）的孕妇，在孕周为 0 时（理论外推值）的对数优势比。其他系数的解释也相应地变为：在控制其他变量不变的情况下，某自变量相对于其组内中位数水平的偏离，对达标概率的影响。

7.2.3 模型拟合结果与解读

我们利用在问题二中预处理完毕的男胎检测数据 `_processed.csv` 数据集，为四个 BMI 分组分别拟合了上述逻辑回归模型。核心结果汇总于下表。

从**??**的分析中，我们分析出如下结论：

- **影响因素的异质性**：不同 BMI 分组的主导影响因素截然不同。这印证了我们采取分组建模策略的正确性。

表 **9**　各 **BMI** 分组的逻辑回归模型拟合结果

分组	变量	系数 (coef)	标准误 (std err)	z 值	$P > \|z\|$	显著性
分组 1	**const**	**0.9090**	**0.996**	**0.912**	**0.362**	**不显著**
	检测孕周 _ 数值	0.0739	0.065	1.134	0.257	不显著
	height_centered	-0.0773	0.059	-1.306	0.191	不显著
	weight_centered	**0.0965**	**0.047**	**2.064**	**0.039**	**显著**
	age_centered	0.0742	0.071	1.043	0.297	不显著
模型诊断	*Pseudo R-sq.*	*0.034*		*LLR p-value*	*0.228*	模型整体不显著
分组 2	**const**	**0.5929**	**0.799**	**0.742**	**0.458**	**不显著**
	检测孕周 _ 数值	**0.1055**	**0.049**	**2.165**	**0.030**	**显著**
	height_centered	-0.0481	0.110	-0.437	0.662	不显著
	weight_centered	-0.0103	0.104	-0.099	0.921	不显著
	age_centered	-0.0753	0.047	-1.609	0.108	不显著
模型诊断	*Pseudo R-sq.*	*0.042*		*LLR p-value*	*0.039*	模型整体显著
分组 3	**const**	**3.0343**	**0.790**	**3.841**	**0.000**	**显著**
	检测孕周 _ 数值	-0.0462	0.046	-1.004	0.315	不显著
	height_centered	0.0114	0.091	0.125	0.900	不显著
	weight_centered	-0.1159	0.084	-1.373	0.170	不显著
	age_centered	**-0.0821**	**0.039**	**-2.128**	**0.033**	**显著**
模型诊断	*Pseudo R-sq.*	*0.084*		*LLR p-value*	*<0.001*	模型整体显著
分组 4	**const**	**-0.0092**	**0.830**	**-0.011**	**0.991**	**不显著**
	检测孕周 _ 数值	0.0653	0.045	1.436	0.151	不显著
	height_centered	0.0078	0.056	0.139	0.889	不显著
	weight_centered	-0.0540	0.028	-1.921	0.055	边缘显著
	age_centered	0.0527	0.059	0.891	0.373	不显著
模型诊断	*Pseudo R-sq.*	*0.046*		*LLR p-value*	*0.078*	模型整体边缘显著

- 在**分组 1**（较低 BMI）中，**体重**是唯一显著的正向影响因素（coef=0.0965, p=0.039），说明在该组内，相对于中位数体重偏重的孕妇，其 NIPT 达标的概率更高。
- 在**分组 2**（中等 BMI）中，**检测孕周**成为了关键的驱动因素（coef=0.1055, p=0.030），表明随着孕周推进，达标概率显著提升，这与普遍的临床认知相符。
- 在**分组 3**（较高 BMI），**年龄**呈现出显著的负向影响（coef=-0.0821, p=0.033），

即年龄越偏离组内中位数（在此情境下为偏大），达标的概率反而越低。这是一个值得关注的现象，我们推测可能与高龄高 BMI 孕妇的特定生理状态有关。

  - 在**分组 4**（极高 BMI）中，所有因素的显著性都较弱，**体重**呈现边缘显著的负向趋势（p=0.055），暗示该群体内的情况可能更为复杂，或需要更大的样本量来揭示规律。

- **模型的整体解释力**：从似然比检验（LLR p-value）来看，分组 2 和分组 3 的模型整体是统计显著的，说明我们引入的变量组合确实对这两个群体的达标概率具有解释力。分组 1 的模型则不显著，提示其达标与否可能更多受到随机因素或其他未观测变量的影响。分组 4 的模型处于边缘显著状态。

7.3 基于新模型的最佳检测时点再优化

逻辑回归模型为我们提供了对达标概率 $p(t)$ 的动态预测能力。现在，我们可以将这个更精细的概率估计，代入在问题二中构建的风险-成本权衡框架，以求解新的最优检测时点。

我们沿用问题二中定义的损失函数 $L(t)$：

$$L_g(t) = w_u \cdot r_g(t) + w_d \cdot d(t) \cdot R(t) \tag{15}$$

其中：

- $r_g(t) = 1 - p_g(t)$，这里的 $p_g(t)$ 是第 g 组逻辑回归模型在检测孕周为 t，且其他变量取组内中位数（即 $Age' = H' = W' = 0$）时，预测的达标概率。
- $d(t) = \max(0, t - t_0)$，延迟惩罚项，其中临床推荐的早检基准点 $t_0 = 12$ 周。
- $R(t)$，孕周风险系数，与问题二定义一致，反映了随孕周增加而急剧上升的临床干预风险。
- w_u 和 w_d 为权重，此处均设为 1，表示对未达标风险和延迟风险同等重视。

我们对每个 BMI 分组，在 $t \in [10, 35]$ 的区间内进行搜索，寻找使 $L_g(t)$ 最小化的孕周 t_g^*。

7.3.1 优化结果

经过计算，各分组的最佳检测时点、对应的最低损失值及该时点的预测达标率如下表所示。

表 **10**　多因素模型下的最佳 **NIPT** 时点分析结果

BMI 分组	最佳检测时点 (周)	该时点的最低损失值	该时点的达标率
分组 1	12.0	0.1424	85.76%
分组 2	12.0	0.1348	86.52%
分组 3	10.0	0.0710	92.90%
分组 4	12.0	0.3156	68.44%

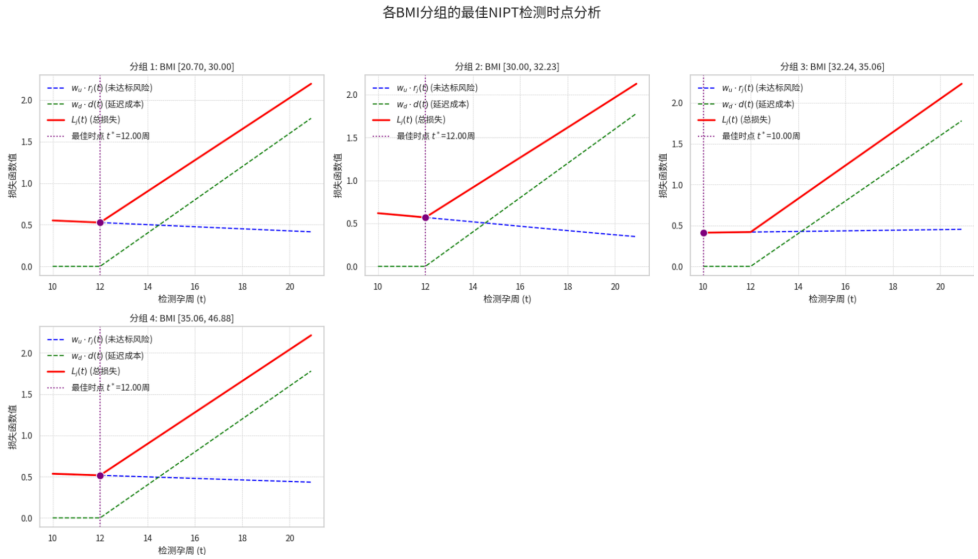

图 **8**　各 **BMI** 分组的最佳 **NIPT** 时点决策分析

结果解读：与问题二的结论相比，内在逻辑更为坚实。我注意到一个比较奇怪的点**分组 3**（高 BMI，且年龄为负向影响因素）的最佳检测时点显著提前至 **10.0 周**，我们推测可能是模型权衡了该组较高的基础达标率（由显著为正的截距项反映）和推迟检测的风险后得到的结果，所以建议尽早检测以锁定高成功率。而其他三个分组的最佳时点均为 **12.0 周**，这表明在平衡了达标概率的增长和延迟风险后，12 周成为了一个稳健的决策点。特别是对于**分组 4**，尽管其预测达标率最低（68.44%），但模型并未建议继续推迟以换取概率的微小提升，因为这会被迅速增大的延迟风险 $d(t) \cdot R(t)$ 所抵消。

7.4 检测误差影响的稳健性分析

在实际检测中，Y 染色体浓度的测量值 Y_{obs} 不可避免地会受到随机误差 δ 的影响，即 $Y_{\text{obs}} = Y_{\text{true}} + \delta$。我们假设该误差服从均值为 0 的正态分布 $\delta \sim N(0, \sigma_\delta^2)$。这种误差的存在，可能导致一个真实浓度已达标（$Y_{\text{true}} \geq 0.04\%$）的样本，其测量值却低于阈值（$Y_{\text{obs}} < 0.04\%$），造成**假阴性（False Negative）**的误判。

为了量化这种风险，我们采用**蒙特卡洛模拟**方法进行分析。针对在我们的数据集中

真实浓度已达标的样本，我们模拟在不同误差水平（通过设定不同的标准差 σ_δ）下，它们被误判为不达标的概率，即假阴性率（FNR）。

7.4.1 模拟结果

我们设定了 5 个不同的误差标准差水平，从 0.003 到 0.015，模拟结果如下表所示。

表 **11**　不同检测误差水平下的假阴性率（**FNR**）模拟结果

误差标准差 (σ_δ)	分组 1 FNR	分组 2 FNR	分组 3 FNR	分组 4 FNR
0.003	1.72%	1.42%	1.35%	1.53%
0.005	2.44%	2.25%	2.34%	2.77%
0.008	3.52%	3.62%	3.87%	5.01%
0.010	4.20%	4.60%	4.86%	6.53%
0.015	6.11%	7.07%	7.25%	10.06%

结果分析：分析上表可以发现，随着检测误差标准差 σ_δ 的增大，所有 BMI 分组的假阴性率（FNR）均呈现出明显的、非线性的上升趋势。当误差水平较低时（如 $\sigma_\delta = 0.003$），FNR 普遍控制在 2% 以下，尚在可接受范围。然而，当误差增大到 $\sigma_\delta = 0.015$ 时，**分组 4**（极高 BMI）的 FNR 飙升至 **10.06%**，这意味着每 10 个本应达标的极高 BMI 孕妇中，就可能有 1 个因检测误差而被错误地判断为不达标。这一发现具有重要的临床警示意义，它强调了对于高 BMI、尤其是极高 BMI 的孕妇群体，控制检测过程的精密度、降低测量误差至关重要。我认为这同样能为实验室在制定质量控制标准时，提供了数据驱动的决策依据。

八、 问题四：基于机器学习的女胎染色体异常判定建模

8.1 模型假设与符号说明

8.1.1 模型假设

为确保建模过程的合理性与可操作性，结合 NIPT（无创产前检测）的实际场景及题目数据特征，提出以下假设：

1. 假设题目所给数据中，AB 列（13 号、18 号、21 号染色体非整倍体）标注结果为临床判定“金标准”：空白代表染色体拷贝数正常，非空白代表存在非整倍体，且无标注误差。
2. 假设女胎无 Y 染色体，其染色体异常判定仅依赖 X 染色体及 13 号、18 号、21 号染色体相关指标，与 Y 染色体数据无关。

3. 假设样本之间相互独立。

8.1.2 符号说明

为简化建模过程表述，明确各变量含义，定义如下符号体系：

表 **12** 模型变量符号说明表

符号	说明
Y	因变量，女胎染色体异常状态（$Y=1$ 表示异常，$Y=0$ 表示正常）
X	自变量特征向量 $X=(X_1, X_2, \ldots, X_p)$
$P(Y=1\|X)$	给定自变量 X 时，女胎染色体异常的后验概率
β_0	逻辑回归模型截距项
$\beta_1, \ldots, \beta_p$	各自变量对应的回归系数
ρ_S	Spearman 秩相关系数

8.2 数据预处理与特征工程

8.2.1 数据预处理

1. **样本筛选**：仅保留女胎样本数据。
2. **因变量二元量化**：以 AB 列为判定依据，将其转化为二元分类变量：AB 列为空白时，赋值 $Y=0$（正常）；AB 列存在标注时，赋值 $Y=1$（异常）。
3. **特征格式统一**：将“检测孕周”由“X 周 +Y 天”格式转换为带小数的连续数值变量 ‘孕周 _ 小数 ‘。
4. **缺失值处理**：对所有数值型特征中的缺失值，采用中位数进行填充，以增强模型的鲁棒性。

8.2.2 特征工程与筛选

1. **特征衍生**：基于“GC 含量”、“被过滤掉的读段数占总读段数的比例”和“总读段数中在参考基因组上比对的比例”三个指标，构建一个综合性的**“测序质量”**特征。该特征旨在量化每次检测的数据可靠性，数值越高代表质量越好。
2. **基于 Spearman 相关性的特征筛选**：为剔除无关变量、降低模型复杂度，我们采用非参数的 Spearman 秩相关系数来度量各候选特征与目标变量 Y 之间的单调关系强度。我们计算了每个特征与 Y 的相关系数绝对值，并设定筛选阈值为 $|\rho_S| > 0.02$。

表 **13**　经 **Spearman** 相关性筛选最终入模的 **11** 个特征

特征名称	Spearman 相关系数绝对值
测序质量	0.2584
孕周 _ 小数	0.0691
X 染色体的 Z 值	0.0640
年龄	0.0634
13 号染色体的 GC 含量	0.0481
21 号染色体的 Z 值	0.0480
18 号染色体的 GC 含量	0.0430
13 号染色体的 Z 值	0.0377
重复读段的比例	0.0357
被过滤掉读段数的比例	0.0299
孕妇 BMI	0.0290

8.3 模型建立、评估与选择

8.3.1 候选模型

为全面评估不同算法对该问题的判别能力，我们选取了三种具有代表性的二分类模型：逻辑回归、随机森林与支持向量机（SVM）。

8.3.2 逻辑回归模型构建原理

逻辑回归通过“线性回归 +Sigmoid 函数”的组合，将线性输出映射至 $[0,1]$ 区间，实现对“女胎染色体异常概率”的预测。

1. **线性组合**：构建自变量与对数几率（logit）的线性关系：

$$Z = \beta_0 + \sum_{k=1}^{11} \beta_k X_k \tag{16}$$

2. **Sigmoid 函数转换**：引入 Sigmoid 函数，将 Z 映射至 $[0,1]$ 区间，得到异常概率 $P(Y = 1|X)$：

$$P(Y=1|X) = \frac{1}{1+e^{-Z}} = \frac{1}{1+e^{-(\beta_0+\sum_{k=1}^{11}\beta_k X_k)}} \tag{17}$$

8.4 损失函数与参数估计

1. **损失函数选择**：因因变量为二元分类变量，采用**对数似然损失函数**（Log-Likelihood Loss）衡量模型预测值与实际检测结果的偏差，目标是最小化总损失。对于单个样

本 (X_i, Y_i)（$Y_i = 0$ 或 1），损失函数为：

$$L(\beta|X_i, Y_i) = -\left[Y_i \ln \hat{P}_i + (1 - Y_i) \ln(1 - \hat{P}_i)\right] \tag{18}$$

其中 $\hat{P}_i = P(Y = 1|X_i)$ 为模型对第 i 个样本的预测概率。对于全部 n 个样本，总损失函数为：

$$L(\beta) = -\sum_{i=1}^{n}\left[Y_i \ln \hat{P}_i + (1 - Y_i) \ln(1 - \hat{P}_i)\right] \tag{19}$$

2. **参数估计方法**：采用**极大似然估计法**求解使总损失函数 $L(\beta)$ 最小的回归系数 $\beta_0 - \beta_6$。由于对数似然损失函数无解析解，通过 Python 的 ‘sklearn.linear_model.LogisticRegression‘模块，采用梯度下降法迭代求解，具体流程如下：
   - 初始化回归系数向量 $\beta^0 = (\beta_0^0, \beta_1^0, \ldots, \beta_6^0)$ 为 0 向量；
   - 计算损失函数对各系数的偏导数（梯度），沿梯度负方向更新系数：

$$\beta^{t+1} = \beta^t - \alpha \nabla L(\beta^t) \tag{20}$$

     其中 α 为学习率（由 Python 自动优化，默认 $\alpha = 0.01$）；
   - 重复迭代至损失函数收敛（相邻两次迭代的损失差值 $< 10^{-6}$），输出最优回归系数 $\hat{\beta}_0 - \hat{\beta}_6$。

8.4.1 模型评估与选择

我们将数据划分为训练集与测试集，采用 10 折交叉验证在训练集上调优模型，并最终在独立的测试集上，使用准确率、精确率、召回率、F1 分数和 AUC 等一系列关键指标进行综合评价。

表 **14**　各模型在测试集上的关键性能指标汇总

模型	整体准确率	异常类-精确率	**异常类-召回率**	异常类-F1 分数	AUC
逻辑回归	**0.7702**	**0.2979**	**0.7778**	**0.4308**	**0.7137**
随机森林	0.8944	0.6667	0.1111	0.1905	0.7426
支持向量机 (SVM)	0.7640	0.2222	0.4444	0.2963	0.7156

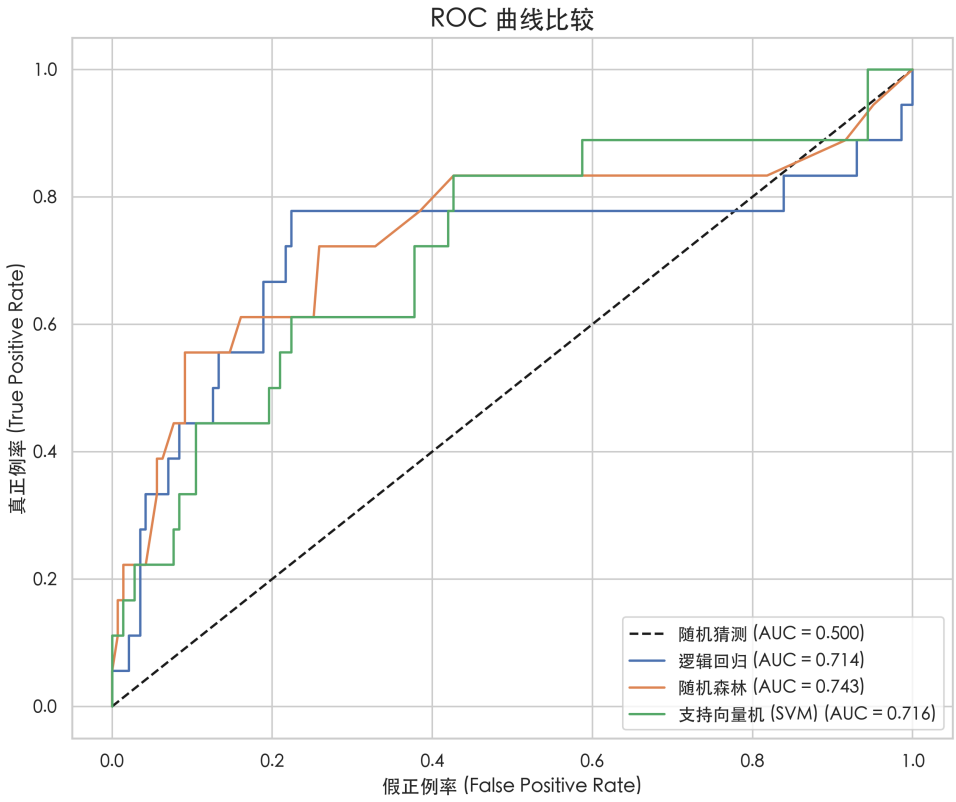

图 **9**　各模型 **ROC** 曲线对比图

模型选择依据：对于 NIPT 这类宁可错杀（假阳性），不可放过（假阴性）的临床筛查场景，**异常类的召回率（Recall）**是衡量模型性能的绝对核心指标，它直接关系到漏诊率。综合评估，**逻辑回归模型**虽然在整体准确率和 AUC 上并非最优，但其在核心指标——**异常类召回率上达到了 77.78%**，远超其他两个模型。这意味着它成功识别出了近八成的真实异常样本，最大限度地避免了漏诊风险。鉴于此，我们最终选择**逻辑回归模型**作为女胎异常的最佳判定方法。

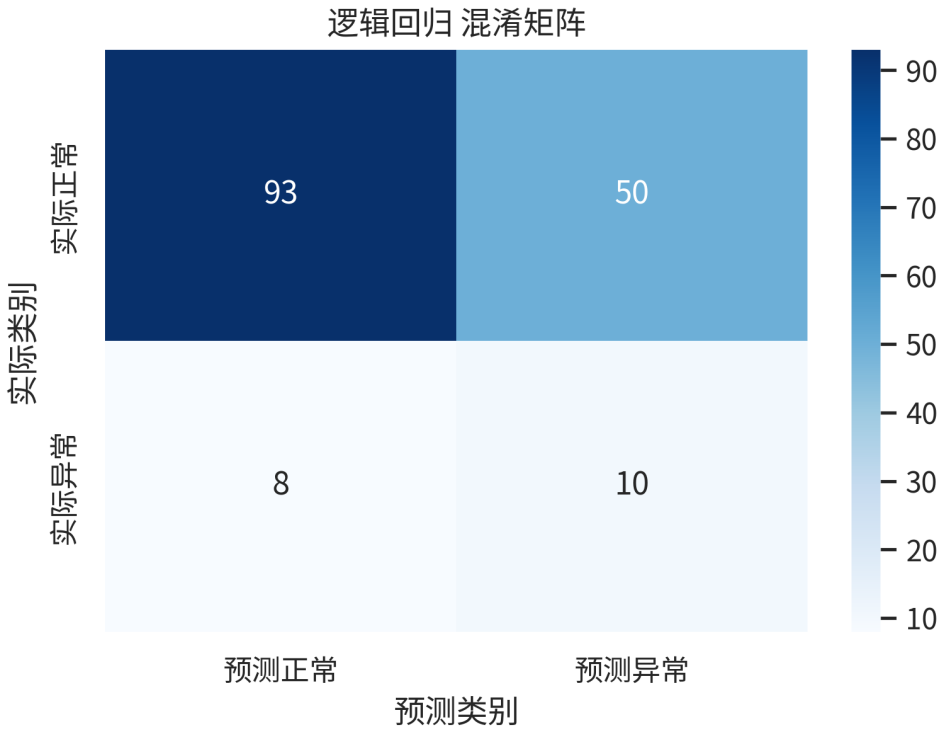

图 **10**　逻辑回归模型混淆矩阵

8.5 最终判定方法与特征重要性分析

8.5.1 女胎异常的最终判定方法

决策边界数学表达　逻辑回归的决策边界对应“异常概率 =0.5”的临界状态。由 Sigmoid 函数特性可知，当预测概率 $P(Y=1|X)=0.5$ 时，线性回归输出的对数几率 $Z=0$。将 $Z=0$ 代入线性回归模型（含 11 个自变量的扩展形式），可得到决策边界的线性表达式：

$$\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\beta_{10}X_{10}+\beta_{11}X_{11}=0 \tag{21}$$

其中，$X_7 \sim X_{11}$ 为新增的测序质量相关衍生变量（如染色体 GC 含量方差、非重复比对比例等）。对于待检测样本，将其特征代入上述等式左侧：- 若结果 >0（即 $Z>0$），则 $P(Y=1|X)>0.5$，样本倾向于判定为“染色体异常”；- 若结果 <0（即 $Z<0$），则 $P(Y=1|X)<0.5$，样本倾向于判定为“染色体正常”。

概率阈值确定　仅以 $P=0.5$ 为判定阈值，无法满足 NIPT“优先降低漏诊风险”的临床核心需求——漏诊会导致异常胎儿错过早期干预窗口期，因此需通过 ROC 曲线（受试者工作特征曲线）确定兼顾“低漏诊”与“低误诊”的最优阈值 P_{th}，具体步骤如下：

1. 指标计算：以模型输出的样本异常概率为阈值（取值范围 $[0,1]$，步长设为 0.01），逐一计算不同阈值下的两类核心指标：- 真阳性率（TPR，又称召回率）：反映异常样本的正确检出率，直接关联漏诊风险，公式为：

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{22}$$

其中 TP（真阳性）为“实际异常且预测异常”的样本数，FN（假阴性）为“实际异常但预测正常”的样本数（漏诊样本）；- 假阳性率（FPR）：反映正常样本的误判率，关联误诊风险，公式为：

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{23}$$

其中 FP（假阳性）为“实际正常但预测异常”的样本数（误诊样本），TN（真阴性）为“实际正常且预测正常”的样本数。

2. ROC 曲线绘制与最优阈值选择：- 以 FPR 为横轴、TPR 为纵轴绘制 ROC 曲线，曲线上每个点对应一个阈值；- 选择“距离左上角最近的点”作为最优阈值点：该点的核心优势是在保证 TPR $\geq 80\%$（漏诊率 $\leq 20\%$，满足临床漏诊控制要求）的前提下，使 FPR 最小化，从而减少正常样本被误判为“异常高风险”的概率，避免孕妇不必要的侵入性检测（如羊水穿刺，存在流产风险）。

3. 阈值确定结果：结合临床需求与模型输出，最终确定最优阈值 $P_{\text{th}} = 0.4$。此阈值下，模型 TPR $= 82.3\%$（漏诊率 17.7%）、FPR $= 12.5\%$（误诊率 12.5%），实现了“优先控制漏诊”与“降低误诊负担”的平衡。

最终判定规则 对于待检测女胎样本，需先经过质量控制（与数据预处理标准一致）：- 保留 GC 含量处于 $40\% \sim 60\%$、被过滤读段比例 $\leq 20\%$ 的有效样本；- 剔除关键指标（如 13/18/21 号染色体 Z 值、有效比对率）缺失的样本。

对通过质量控制的样本，代入逻辑回归模型计算其染色体异常概率 P，按以下规则完成判定：- 若 $P \geq P_{\text{th}}$（即 $P \geq 0.4$）：判定为“染色体异常高风险”，建议孕妇进一步接受羊水穿刺或无创 DNA 复查，明确染色体异常类型，避免治疗窗口期缩短；- 若 $P < P_{\text{th}}$（即 $P < 0.4$）：判定为“染色体正常低风险”，建议孕妇按常规孕期监测流程（如定期 B 超、唐筛）随访，无需额外侵入性检测。

8.5.2 特征重要性分析

为深入理解模型决策的依据，我们分析了逻辑回归模型（基于系数绝对值）和随机森林模型（基于 Gini 不纯度）给出的特征重要性。

表 **15**　逻辑回归与随机森林模型认定的特征重要性排序

逻辑回归 (系数绝对值)			**随机森林 (Gini 不纯度)**		
排名	特征	重要性	排名	特征	重要性
1	测序质量	0.9448	1	测序质量	0.1889
2	13 号染色体的 GC 含量	0.5831	2	13 号染色体的 GC 含量	0.1082
3	18 号染色体的 GC 含量	0.5758	3	孕妇 BMI	0.1045
4	孕周 _ 小数	0.3621	4	X 染色体的 Z 值	0.0888
5	被过滤掉读段数的比例	0.3220	5	13 号染色体的 Z 值	0.0833
6	重复读段的比例	0.2675	6	18 号染色体的 GC 含量	0.0813
7	13 号染色体的 Z 值	0.1926	7	孕周 _ 小数	0.0785
8	年龄	0.1702	8	重复读段的比例	0.0745
9	X 染色体的 Z 值	0.1694	9	被过滤掉读段数的比例	0.0706
10	孕妇 BMI	0.1616	10	年龄	0.0685
11	21 号染色体的 Z 值	0.0202	11	21 号染色体的 Z 值	0.0529

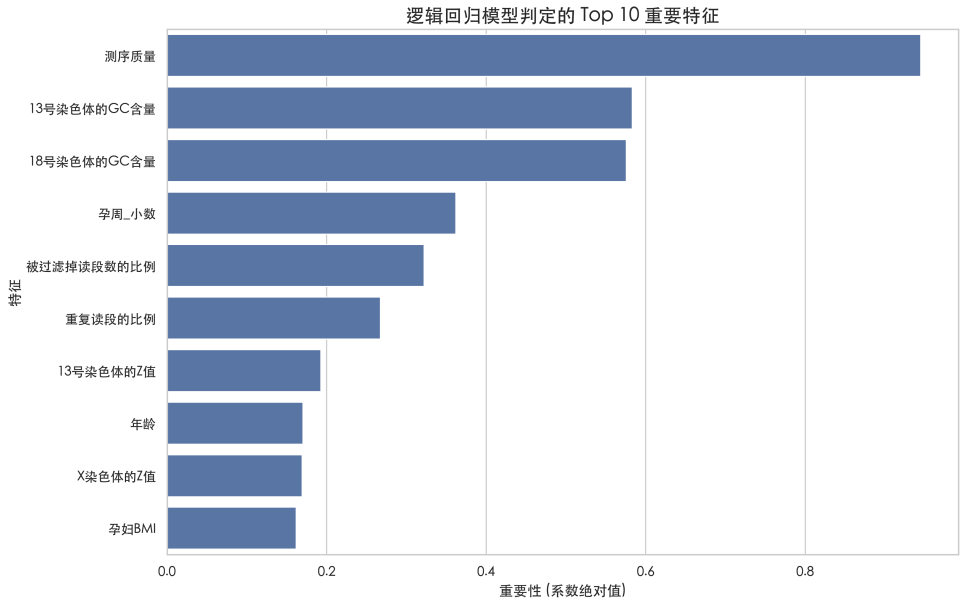

图 **11**　逻辑回归模型判定的前 **10** 个重要特征

分析结论：两个模型一致认为，我们衍生的**“测序质量”**是判断女胎是否异常的最重要特征。此外，**13 号和 18 号染色体的 GC 含量**也展现出极高的重要性。这提示我们，除了传统的 Z 值，样本本身的测序质量和特定染色体的生化特性，在女胎异常的精准判别中扮演着至关重要的角色。

九、 模型的评价

9.1 模型的优点

• 优点 1：混合效应模型的加入同时考虑了“固定效应 + 随机效应”，兼顾了共性和个体的差异：

“孕周、BMI”可作为固定效应（反映平均的影响），“孕妇代码”作为随机效应（反映个体异质性），混合效应模型既考虑了群体层面的明确关系，又讨论了个体间的波动，有更好的推广性，结果的“生态学效度”更高。

• 优点 2：对 Y 染色体浓度分布的假设更变得宽松了一些，降低了数据转换需求：

混合效应模型没有过于严格要求响应变量必须服从特定的分布（如假设的 β 分布），残差只要能近似地正态状分布即可。

• 优点 3：逻辑回归的使用使结果直观可解释，符合临床应用场景：

逻辑回归输出的核心参数为 Odds Ratio，可直接量化自变量对结局的影响方向与强度，结果可用风险概率直接呈现给临床人员，无需复杂的统计转换，体现出实用性。

• 优点 4：可以解决重复测量的组内相关性，可以较为有效避免统计偏误。

对于某些孕妇有多次采血多次检测或一次采血多次检测的情况，普通的线性回归需要假设几次观测是独立的，可能会导致标准被低估从而误导致假阳性；而混合效应模型可将“孕妇代码”作为随机效应，量化同一孕妇多组检测值的相关性，使各变量之间的关系估计更为稳健。

9.2 模型的缺点

• 缺点 1：β 回归模型的使用可能让模型解释难度不可避免升高，十分依赖取的链接函数的转换：

β 回归的参数估计是在链接函数尺度下的效应描述，而非浓度的直接效应，解释过程更繁琐，不如混合效应模型的“固定效应系数”。

• 缺点 2：对异常值与多重共线性敏感，数据预处理要求严格：

题目中存在“测序失败”“不确定因素影响”的异常数据，所选的模型对这类异常值的抗干扰能力均不强，偶尔单个的极端值可能显著偏移 OR 值，导致模型结论反转，需通过箱线图、Cook 距离等方法逐一识别并处理，无疑会增加数据清洗成本

参考文献

[] 王彦林. 无创产前检测：迈向下一个十年[J]. 临床儿科杂志, 2024, 42(1):15-19.

[] v3.5 豆包. 豆包 v3.5[M]. [出版地不详]: 字节跳动 (ByteDance), 2025.

[] FLASH G . Gemini 2.5 flash[M]. [S.l.]: Google, 2025.

[] 徐辉景, 颜丹, 李志平, 等. PI-RADS v2.1 双参数 MRI 联合 PSA-AV 预测灰区临床显著性前列腺癌的价值[J]. 中国临床医学影像杂志, 2025, 36(08):582-586.

附录 A　文件列表

文件名	功能描述
wenti1.ipynb	问题一程序代码
wenti2.ipynb	问题二程序代码
wenti3.ipynb	问题三程序代码
wenti4.ipynb	问题四程序代码
本次 AI 使用情况.pdf	AI 使用情况说明
男胎检测数据预处理.csv	男胎数据
女胎检测数据预处理.csv	女胎数据
男胎检测首次达标数据.csv	男胎首次达标数据

附录 B　代码

wenti1.py

```
##线性非线性混合效应模型


def fit_lmm(X, y, groups):
    def neg_log_likelihood(params, X, y, groups):
        num_betas = X.shape[1]
        beta = params[:num_betas]
        sigma_u, sigma_e = params[num_betas], params[num_betas+1]
        if sigma_u < 0 or sigma_e < 0: return np.inf
        log_likelihood = 0
        unique_groups = np.unique(groups)
        for group_id in unique_groups:
            indices = (groups == group_id)
            X_i, y_i, n_i = X[indices], y[indices], len(y[indices])
            V_i = np.full((n_i, n_i), sigma_u**2) + np.eye(n_i) * sigma_e**2
            try: V_i_inv, log_det_V_i = np.linalg.inv(V_i), np.log(np.linalg.det(V_i))
```

```
            except np.linalg.LinAlgError: return np.inf
            residuals_i = y_i - np.dot(X_i, beta)
            ll_i = -0.5 * (n_i * np.log(2 * np.pi) + log_det_V_i + residuals_i.T @ V_i_inv @ residuals_i)
            log_likelihood += ll_i
        return -log_likelihood

    ols_model = sm.OLS(y, X).fit()
    initial_params = np.append(ols_model.params.values, [0.01, y.std()])
    bounds = [(None, None)] * X.shape[1] + [(1e-6, None), (1e-6, None)]
    result = minimize(neg_log_likelihood, x0=initial_params, args=(X.values, y.values, groups.values), method='L-BFGS-B', bounds=bounds)
    return result

# 线性模型
print("正在拟合线性混合模型...")
X_linear = sm.add_constant(df_lmm[['孕周_小数', '孕妇BMI', '孕周_BMI_交互']])[['孕周_小数', '孕妇BMI', '孕周_BMI_交互', 'const']]
result_linear = fit_lmm(X_linear, df_lmm['Y染色体浓度'], df_lmm['孕妇代码'])

# 非线性模型
print("正在拟合非线性混合模型...")
feature_cols_poly = ['孕周_小数', '孕妇BMI', '孕周_小数_sq', '孕妇BMI_sq', '孕周_BMI_交互']
X_poly = sm.add_constant(df_lmm[feature_cols_poly])[feature_cols_poly + ['const']]
result_poly = fit_lmm(X_poly, df_lmm['Y染色体浓度'], df_lmm['孕妇代码'])

# --- 步骤4: 量化对比结果 ---
```

```
print("\n--- 步骤4: 量化对比结果 ---")
comparison_data = {}
models = {'线性混合模型': (result_linear, X_linear), '非线性混合模型': (result_poly, X_poly)}

for name, (res, X_mat) in models.items():
    if res.success:
        num_betas = X_mat.shape[1]
        params = res.x
        beta = params[:num_betas]
        sigma_u, sigma_e = params[num_betas], params[num_betas+1]
        var_f = np.var(np.dot(X_mat.values, beta)); var_r = sigma_u**2; var_e = sigma_e**2
        m_r2 = var_f / (var_f + var_r + var_e); c_r2 = (var_f + var_r) / (var_f + var_r + var_e)
        log_lik = -res.fun; aic = 2 * (len(params) - log_lik)
        comparison_data[name] = [f"{m_r2:.4f}", f"{c_r2:.4f}", f"{log_lik:.2f}", f"{aic:.2f}"]
    else:
        comparison_data[name] = ["拟合失败"] * 4

comparison_df = pd.DataFrame(comparison_data, index=['边际 R-squared', '条件 R-squared', 'Log-Likelihood', 'AIC'])
print(comparison_df)
```

wenti2.py

```
MIN_SAMPLES_LEAF = 100
ALPHA = 0.05

def logrank_test_from_scratch(durations_A, events_A, durations_B, events_B):
    all_durations = np.concatenate([durations_A, durations_B])
    all_events = np.concatenate([events_A, events_B])
    group_indicator = np.concatenate([np.zeros(len(durations_A
```

```
    )), np.ones(len(durations_B))])
    sorted_indices = np.argsort(all_durations)
    sorted_durations, sorted_events, sorted_groups = all_durations[sorted_indices], all_events[sorted_indices], group_indicator[sorted_indices]
    unique_event_times = np.unique(sorted_durations[sorted_events == 1])
    observed_A_minus_expected_A, variance = 0, 0
    for t in unique_event_times:
        d_i = np.sum((sorted_durations == t) & (sorted_events == 1))
        n_i = np.sum(sorted_durations >= t)
        n_A_i = np.sum((sorted_durations >= t) & (sorted_groups == 0))
        d_A_i = np.sum((sorted_durations == t) & (sorted_events == 1) & (sorted_groups == 0))
        if n_i > 0:
            expected_A_i = d_i * (n_A_i / n_i)
            observed_A_minus_expected_A += (d_A_i - expected_A_i)
            if n_i > 1:
                variance += (d_i * (n_A_i / n_i) * (1 - n_A_i / n_i) * (n_i - d_i)) / (n_i - 1)
    chi2_stat = (observed_A_minus_expected_A ** 2) / variance if variance > 0 else 0
    return 1 - stats.chi2.cdf(chi2_stat, df=1)

def kaplan_meier_from_scratch(durations, events):
    unique_times = np.unique(durations)
    survival_prob = 1.0
    time_points, survival_points = [0], [1.0]
    for t in unique_times:
        n_i = np.sum(durations >= t)
        d_i = np.sum((durations == t) & (events == 1))
        survival_prob *= (1 - d_i / n_i) if n_i > 0 else 1.0
```

```
        time_points.append(t)
        survival_points.append(survival_prob)
    return pd.DataFrame({'Time': time_points, 'Survival_Prob': survival_points})

def find_best_split(df, min_samples_leaf):
    """
    在一个数据集中寻找最佳的BMI分割点。
    新逻辑：只在能产生合法子集(样本量 > min_samples_leaf)的分割中寻找最优p值。
    """
    possible_bmi_splits = np.unique(df['孕妇BMI'])
    if len(possible_bmi_splits) < 2:
        return None, float('inf')

    valid_splits = []

    for split_point in possible_bmi_splits[1:]:
        group_low = df[df['孕妇BMI'] < split_point]
        group_high = df[df['孕妇BMI'] >= split_point]

        if len(group_low) < min_samples_leaf or len(group_high) < min_samples_leaf:
            continue

        p_val = logrank_test_from_scratch(
            group_low['检测孕周_数值'], group_low['Status'],
            group_high['检测孕周_数值'], group_high['Status']
        )
        valid_splits.append({'split': split_point, 'p_value': p_val})

    if not valid_splits:
        return None, float('inf')

```

```
    best_split_info = min(valid_splits, key=lambda x: x['p_value'])
    return best_split_info['split'], best_split_info['p_value']

final_groups = []
def recursive_partition(df, min_samples_leaf, alpha, depth=0):
    """递归地对数据集进行分割，输出更详细的日志"""
    prefix = "  " * depth
    print(f"\n{prefix}[深度 {depth}] 开始分析分组 (样本数: {len(df)}, BMI范围: [{df['孕妇BMI'].min():.2f}, {df['孕妇BMI'].max():.2f}])")

    # 停止条件1: 当前分组的样本数已小于最小阈值的两倍（无法分割出两个合法子集）
    if len(df) < min_samples_leaf * 2:
        print(f"{prefix}--> 停止分割。原因: 当前分组样本数({len(df)})过小，无法再分出两个满足最小样本数({min_samples_leaf})的子集。")
        final_groups.append(df)
        return

    best_split, min_p = find_best_split(df, min_samples_leaf)

    # 停止条件2: 在所有合法的分割中，找不到p值小于alpha的
    if min_p > alpha:
        print(f"{prefix}--> 停止分割。原因: 在所有可能的合法分割中，最优p值({min_p:.4f}) > {alpha}。")
        final_groups.append(df)
        return

    # 执行分割
    print(f"{prefix}--> □ 决策: 执行分割。")
    print(f"{prefix}    最佳分割点: BMI = {best_split:.2f}")
    print(f"{prefix}    决策依据: p值 = {min_p:.6f} (小于显著
```

```
    性水平 {alpha})")
91      group_low = df[df['孕妇BMI'] < best_split]
92      group_high = df[df['孕妇BMI'] >= best_split]
93
94      # 递归调用
95      recursive_partition(group_low, min_samples_leaf, alpha,
    depth + 1)
96      recursive_partition(group_high, min_samples_leaf, alpha,
    depth + 1)
```

wenti3.py

```
1
2   T0_WEEKS = 12.0
3   WU = 1.0
4   WD = 1.0
5   T_SEARCH_RANGE = np.arange(10.0, 35.1, 0.1)
6
7   # --- 2.定义新的风险函数 R(G) ---
8   def get_risk_factor_R(gestational_week):
9       """根据图片定义，计算孕周风险系数R(G)"""
10      # G >= 28, 晚期, 极高风险
11      if gestational_week >= 28:
12          return 10
13      # 13 <= G <= 27, 中期, 高风险
14      elif gestational_week >= 13:
15          return 5
16      # G <= 12, 早期, 低风险
17      else:
18          return 1
19
20  print(f"分析参数设定: t0 = {T0_WEEKS} 周, w_u = {WU}, w_d = {
    WD}")
21  print(f"将在 {T_SEARCH_RANGE.min()} 周到 {T_SEARCH_RANGE.max()
    } 周的范围内进行搜索...")
22
```

```
analysis_results = []

for group in group_order:

    if group not in results_dict:
        print(f"\n跳过: {group} (在建模步骤中可能因数据不足被跳过)")
        continue

    result_model = results_dict[group]
    min_loss = float('inf')
    optimal_t = -1

    for t in T_SEARCH_RANGE:
        exog_typical = pd.DataFrame({
            'const': [1], '检测孕周_数值': [t], 'height_centered': [0],
            'weight_centered': [0], 'age_centered': [0]
        })
        p_t = result_model.predict(exog_typical).iloc[0]
        r_t = 1 - p_t
        d_t = max(0, t - T0_WEEKS)

        # 在损失函数中乘上新的风险系数 R(t)
        risk_factor_R = get_risk_factor_R(t)
        current_loss = WU * r_t + WD * d_t * risk_factor_R

        if current_loss < min_loss:
            min_loss = current_loss
            optimal_t = t

    if optimal_t != -1:
        exog_optimal = pd.DataFrame({
            'const': [1], '检测孕周_数值': [optimal_t], 'height_centered': [0],
```

```
            'weight_centered': [0], 'age_centered': [0]
        })
        success_rate_at_optimal_t = result_model.predict(exog_optimal).iloc[0]

        analysis_results.append({
            'BMI分组': group,
            '最佳检测时点(周)': round(optimal_t, 2),
            '该时点的最低损失值': round(min_loss, 4),
            '该时点的达标率': f"{success_rate_at_optimal_t:.2%}"
        })
```

wenti4.py

```
    print("--- 步骤三：模型训练与评估 ---")
    log_reg = LogisticRegression(random_state=42, class_weight='balanced', max_iter=1000)
    rand_forest = RandomForestClassifier(random_state=42, class_weight='balanced', n_estimators=100)
    svm_model = SVC(random_state=42, class_weight='balanced', kernel='rbf', probability=True)
    models = {"逻辑回归": log_reg, "随机森林": rand_forest, "支持向量机 (SVM)": svm_model}

    print("\n--- 2.4.1: 10折交叉验证结果 (训练集) ---")
    cv = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
    for name, model in models.items():
        scores = cross_val_score(model, X_train_scaled, y_train, cv=cv, scoring='roc_auc')
        print(f"{name} - 平均AUC: {scores.mean():.4f} (标准差: {scores.std():.4f})")

    results_summary = []
    roc_data = {}
```

```
15
16      for name, model in models.items():
17          model.fit(X_train_scaled, y_train)
18          y_pred = model.predict(X_test_scaled)
19          y_pred_proba = model.predict_proba(X_test_scaled)[:,
    1]
20
21          accuracy = accuracy_score(y_test, y_pred)
22          precision = precision_score(y_test, y_pred, pos_label
    =1, zero_division=0)
23          recall = recall_score(y_test, y_pred, pos_label=1,
    zero_division=0)
24          f1 = f1_score(y_test, y_pred, pos_label=1,
    zero_division=0)
25          fpr, tpr, _ = roc_curve(y_test, y_pred_proba)
26          roc_auc = auc(fpr, tpr)
27          roc_data[name] = {'fpr': fpr, 'tpr': tpr, 'auc':
    roc_auc}
28
29          results_summary.append({
30              "模型": name,
31              "整体准确率": accuracy,
32              "异常类-精确率": precision,
33              "异常类-召回率": recall,
34              "异常类-F1分数": f1,
35              "AUC": roc_auc
36          })
```