基于广义加性模型与多目标遗传优化的产前检测策略研究

摘要

孕妇的产前检测对于胎儿健康和社会生育率保障具有重要意义，优化产前检测的质量与准确性也因此尤为重要。本研究针对无创产前基因检测（NIPT）最佳检测时点确定的挑战，基于历史孕妇 NIPT 数据，结合孕妇身体相关指标，构建了一系列数学模型，得到了确定最佳 NIPT 时点及检测胎儿异常的方案，旨在为优生优育提供科学有效的指导。

对于问题一，本文首先对整体数据进行预处理和可视化分析，为构建**胎儿 Y 染色体浓度关系模型**提供支撑；之后，结合**分位数归一化方法**与 Spearman **相关系数分析**，确定将**孕周、BMI 与年龄**作为**广义加性模型**（GAM）的输入，使用**张量积平滑方法**刻画变量之间的交互作用，并通过 **B 样条基函数**拟合非线性趋势。随后结合 **KS 检验**，指定 **Gamma 分布**，使用**惩罚似然估计和限制极大似然**(REML)**方法**对模型进行 python 迭代求解，得到拟合成功率为 60.05%，伪 R^2值为 0.5123 的 GAM 模型，并使用三维散点图对结果进行可视化。最后，通过 **F 检验**得出模型各平滑项 p-value，验证了孕周、BMI 和年龄对 Y 染色体浓度确具有显著影响。

对于问题二，为了最小化孕妇面临的综合风险，本文建立了关于 Y 染色体浓度最早达标时间的**假阴性风险**与**时间窗口风险**的**评估函数**，将 BMI 分组、检测时点等设为决策变量，建立了**混合整数线性规划模型。**结合问题一中 **GAM 模型**的拟合结果，使用**双基因变长实数编码**输入**遗传算法**进行问题求解，并绘制了 BMI 分组结果、最佳 NIPT 时点与实际孕妇数据的对比图。结果显示应将孕妇 **BMI 分为 4 个区间：[26.0,34.5)，[34.5,38.2)，[38.2,41.2)，[41.2,47.0)**，且每个区间的**最佳时点为 14，14，16，17 周**。最后，通过**高斯噪声分析法**验证了模型**对于检测误差的鲁棒性**。

对于问题三，本文在问题二的基础上扩展了模型，在对变量进行**多重共线性分析**后，将身高、年龄、检测误差等额外影响因素纳入模型不同阶段，建立了**增强型多目标优化模型**。该模型同样采用改进遗传算法求解。结果表明应将孕妇 BMI **分为 3 个区间：[26.0,39.5)，[39.5,43.7)，[43.7,47.0)**，且每个区间的**最佳时点为 14，15，19 周**。模型同样通过**高斯噪声分析法**验证了对于检测误差的抗干扰性。

对于问题四，本文提出将女胎诊断问题转化为一个**二分类模式识别**问题，通过 Cramer's V 和**点二列相关系数**分析选取重要特征，并结合 SMOTE-Tomek 方法对**逻辑回归分类模型**进行动态调参。结果显示**模型的 AUC 达到** 0.86，具有良好的分类能力。

关键词： 广义加性模型 遗传算法 多目标优化 逻辑回归

一、问题重述

NIPT 是一种通过采集母体血液、检测胎儿的游离 DNA 片段、分析胎儿染色体是否存在异常的产前检测技术,目的是通过早期检测确定胎儿的健康状况。现有研究表明, NIPT 检测精度在很大程度上依赖于胎儿性染色体 DNA 片段在母体血液中的浓度水平。对于男胎孕妇而言，Y 染色体浓度是评估检测准确性的关键指标，当其浓度水平达到 4%以上时，检测结果的准确性能够得到有效保障。然而胎儿 DNA 浓度受到多种生理因素影响，包括孕周、孕妇 BMI、年龄等指标，这些因素相互作用使得最佳检测时机和胎儿是否健康的诊断成为临床实践中的重要挑战。

问题一：建立 Y 染色体浓度和孕妇孕周数、BMI 等指标的相关模型，并做显著性检验。

问题二：男胎孕妇的 BMI 对 Y 染色体浓度的最早达标时间起主要作用，建立模型对男胎孕妇 BMI 进行合理分组，给出每组 BMI 对应的最佳 NIPT 时间点，使得孕妇潜在风险最小，并对检测误差对结果造成的影响进行分析。

问题三：在考虑身高、体重、年龄等多元影响因素基础上，结合检测技术误差和 Y 染色体达标率等实际约束条件，建立更加全面的优化模型，实现基于 BMI 分组的个体化检测方案设计。

问题四： 由于女胎缺乏 Y 染色体这一生物学特点，以染色体非整倍体检测为检测标准，整合 X 染色体指标、各目标染色体的统计学 Z 值、测序质量参数以及孕妇生理指标，构建女胎异常的综合诊断算法。

二、模型假设

假设 1：孕妇的 BMI、年龄、孕周等基本信息准确可靠，数据采集过程中不存在系统性偏差。

假设 2：仅考虑题目所列因素对 Y 染色体浓度的影响，忽略其他可能的次要因素如遗传背景、生活方式、药物使用等对检测结果的影响。

假设 3：在研究期间内，孕妇群体的 BMI 分布特征、年龄结构等人口学特征保

持相对稳定，不发生显著变化。

假设 4：孕妇会遵循医生建议在指定的孕周进行 NIPT 检测，不存在因个人原因导致的检测时点偏移。

假设 5：早期发现胎儿异常比晚期发现具有更低的治疗风险和更多的干预选择，时间窗口风险随检测时点推迟而递增。

假设 6：同一 BMI 组内的孕妇具有相似的生理特征，采用统一的检测时点是合理且可行的。

假设 7：各种因素如环境、技术导致的检测误差为正态分布。

三、符号说明

符号	含义	单位
i	第i个孕妇	-
w	孕周数	周
B_i	孕妇i的 BMI	kg/m^2
A_i	孕妇i的年龄	岁
P_i	孕妇i的怀孕次数	次
H_i	孕妇i的身高	m
w_i^*	标准化后孕妇i的孕周数	周
B_i^*	分位数归一化后孕妇i的 BMI	kg/m^2
$Z_{i,g}$	孕妇i是否被分配到第g组	-
$U_{g,w}$	第g组的检测时点是否为第 w 周	-
$S_{i,w}$	孕妇 i 是否在第 w 周进行检测	-
FN_i	孕妇i检测结果为假阴性概率	-
$R_{i,w*}$	孕妇i在w*周检测出问题的风险系数	-

G	BMI 分组数量	-

四、问题分析

问题一需要分析 Y 染色体浓度作为因变量与孕周、BMI、年龄等多个自变量之间的定量关系，并通过统计检验验证模型的有效性。由于生物医学数据通常具有非线性特征和复杂的交互效应，因此采用多元非线性回归模型建模。我们选择广义加性模型（GAM），该模型能够灵活处理变量间的非线性关系，同时具有平滑非线性趋势的生物医学数据。我们对数据进行数据预处理，包括异常值检测、数据标准化和分位数归一化；然后通过相关性分析筛选核心预测变量；接着构建 GAM 模型，采用 B 样条基函数拟合非线性关系，使用张量积平滑捕捉变量间交互效应；最后通过 F 检验验证模型各项的统计显著性，设置 95%置信区间评估预测精度。

问题二需要在保证检测准确性的前提下寻找最优的 BMI 分组策略和检测时点配置，从而最小化孕妇在产前检测过程中面临的综合风险。综合风险主要包括假阴性风险与时间窗口风险。因此，该问题本质上是一个兼顾检测准确性与检测时效性的多目标优化问题。我们建立了一个混合整数非线性规划模型。模型中既包含连续决策变量也包含离散决策变量，目标函数为基于 GAM 预测的非线性风险函数。由于问题复杂性，我们采用遗传算法求解，设计实数编码方案表示 BMI 切点和检测周，通过轮盘赌选择、高斯变异等遗传操作寻找最优解；最后通过敏感性分析评估检测误差对结果的影响。

问题三是在问题二基础上还要综合检测误差和 Y 染色体达标率约束，使模型更贴近临床实际但复杂度显著提升。为此我们构建增强型 GAM 模型以纳入多元影响因素，建立 Y 染色体达标率评估机制和 GC 含量质量评分体系，形成多层次耦合的优化体系。在模型求解上，通过设计包含噪声项的假阴性风险函数和 BMI 调整的时间风险函数来构建综合目标函数，采用改进遗传算法进行求解，其中增加自适应变异概率和质量约束条件以提高算法性能，最终通过多次独立运行验证解的稳定性和收敛性，确保优化结果的可靠性和实用性。

问题四是一个有监督的二分类问题，需要基于多指标特征融合判定女胎是否存在染色体异常。我们首先使用 Cramer's V 和点二列相关系数分析各个特征对染色体是否异常的重要性，并据此对特征进行选取。然后构建了基于 SMOTETomek 过采样技术的逻辑回归模型，并进行网格化调参。在调参时着重关注了非整倍体的召回率以降低漏诊的危害，最终选取了最优的正则化方式和强度，避免模型因特征过多而过拟合。最后通过 ROC 曲线、召回率和准确率评估了模型的整体分类能力。

五、模型的建立与求解

5.1 问题一 分析胎儿 Y 染色体浓度与孕妇相关指标的相关性

5.1.1 数据预处理

在对影响胎儿 y 染色体浓度的指标进行分析之前，我们先进行数据预处理，这一步骤对于确保分析结果的准确性和可靠性有重要作用。

首先，针对数据格式不统一的问题，我们进行了标准化处理。原始数据中孕周以"周+天"格式记录，我们将其统一转换为连续变量天数。同时，将胎儿健康状态进行二值化编码，健康记为 1，不健康记为 0。

其次，为了直观地看到数据之间的关系，我们对每个指标用直方图进行可视化分析，并且用 K-S 检验对其进行正态性检验。

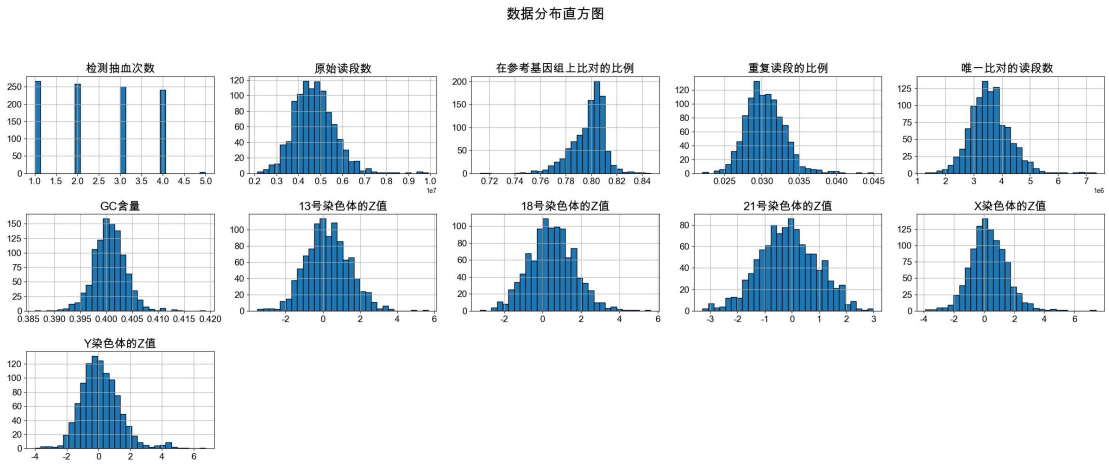


图 1：数据分布直方图

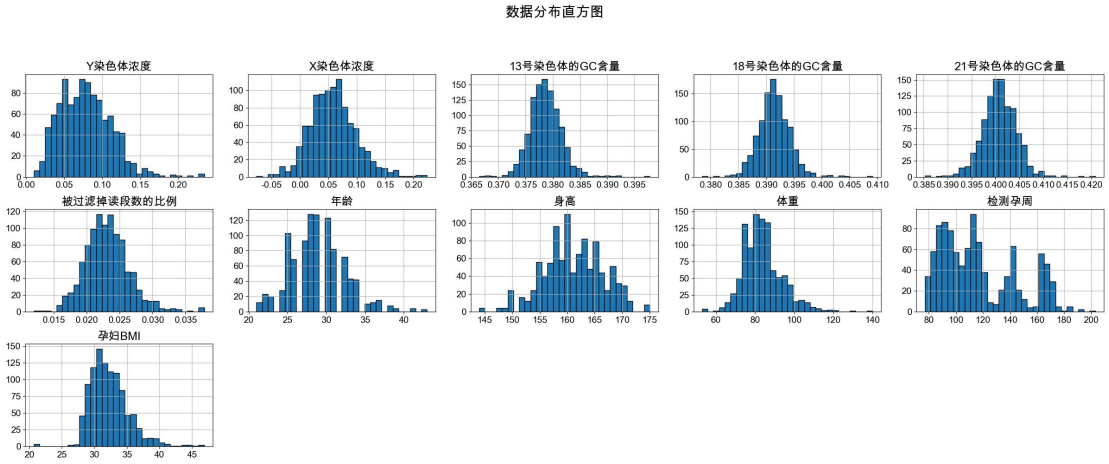


图 2：数据分布直方图

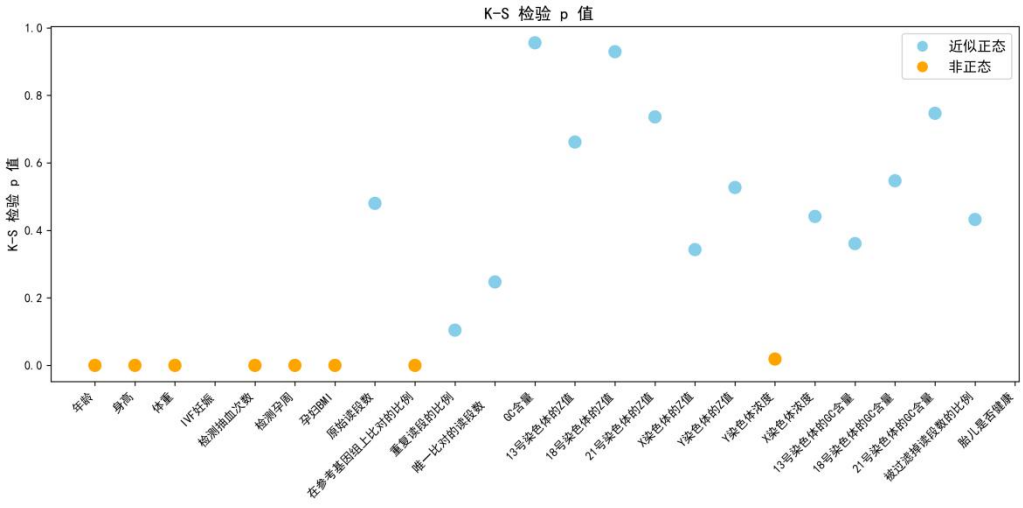


图 3：K-S 检验 p 值

为了避免极端的检测数据对模型准确性造成影响，我们对除了年龄、身高、体重、BMI、检测孕周的检测指标进行了异常值处理。对于连续变量，我们使用四分位距（IQR）方法进行异常值检测。首先计算其第 1 四分位数 Q1 和第 3 四分位数 Q3，得到四分位距 $IQR = Q3 - Q1$。然后将满足 $X < Q1 - 1.5 \times Q3$ 或 $X > Q3 + 1.5 \times IQR$ 的观测值视为异常值并予以删除。

接着，考虑到 BMI 数据呈现右偏分布的特征，偏态分布会导致 GAM 模型的残差不满足正态性假设，影响参数估计的无偏性和有效性，同时会产生异方差问题，为此我们采用分位数归一化方法对 BMI 进行变换。具体方法是将 BMI 的经验分布函数值通过标准正态分布的逆函数进行转换，公式为：

$$B_i^* = \Phi^{-1}(F_B(B_i))$$

其中$F_B(x) = \frac{1}{n}\sum_{i=1}^{n} I(B_i \le x)$为 BMI 的经验分布函数，$I(\cdot)$为示性函数，$\Phi^{-1}(\cdot)$为标准正态分布的逆函数：

$$\Phi^{-1}(p) = \inf\{x \in \mathbb{R} : \Phi(x) \ge p\}$$

其中$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$为标准正态分布的累积分布函数。这样原本的偏态 BMI 分布就转换为近似正态分布$B_i^* \sim N(0,1)$。

我们还对孕周数据进行 Z-score 标准化处理，公式为

$$w_i^* = \frac{w_i - \bar{w}}{s_w}$$

其中w_i为原始孕周天数，$\bar{w}$和s_w分别为样本均值和标准差。这有助于消除量纲的影响，为后续与其他变量进行比较和建模提供便利。

5.1.2 变量筛选与特征选择

鉴于部分特征不符合正态分布，我们采用 Spearman 相关系数来评估各潜在预测变量与 Y 染色体浓度之间的相关性，从而衡量变量的重要性。

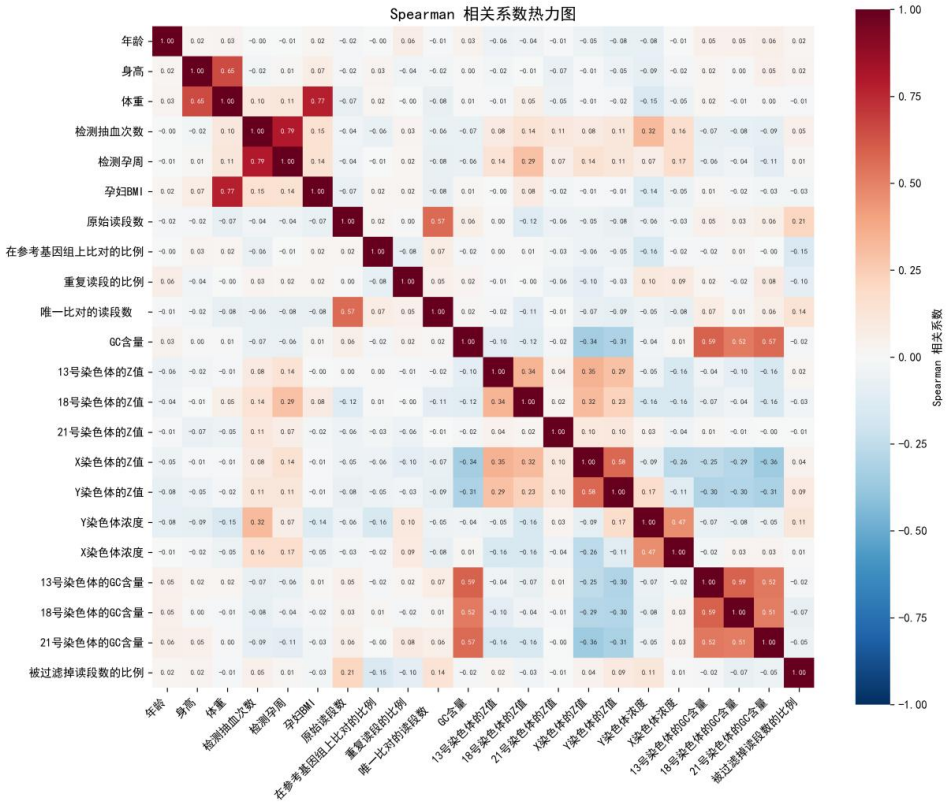


图 4：斯皮尔曼相关系数热力图

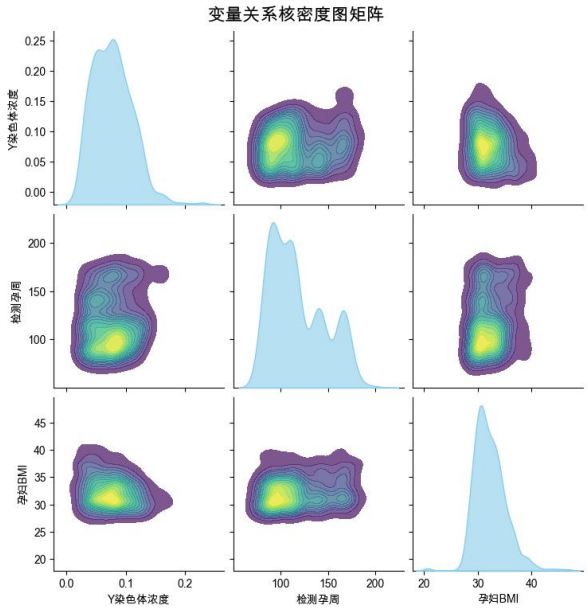


图 5：变量关系核密度图矩阵

对相关性热力图和散点图矩阵进行分析，我们发现 Y 染色体浓度与孕周、BMI、年龄均存在有限的单调相关性。相关性热力图显示 Y 染色体浓度与孕周呈正相关（r≈0.07），与 BMI 呈负相关（r≈-0.14），与年龄呈负相关（r≈-0.08）。

密度图进一步揭示 Y 染色体浓度与孕周、BMI、年龄均呈现明显的非线性分布模式，数据点呈曲线分布而非直线。

我们发现虽然整体样本的孕周和 Y 染色体浓度的单调相关性很弱，但在个体层面表现出了很强的相关性。在排除了检查抽血次数小于三次的孕妇样本后，我们通过绘制 Spearman 相关系数分布图对个体的 Y 染色体浓度与孕周进行分析。可以发现，大部分个体的孕周和 Y 染色体浓度呈现了极高的正相关。这可能是因为生物数据的个体差异明显，导致群体的相关性不显著。

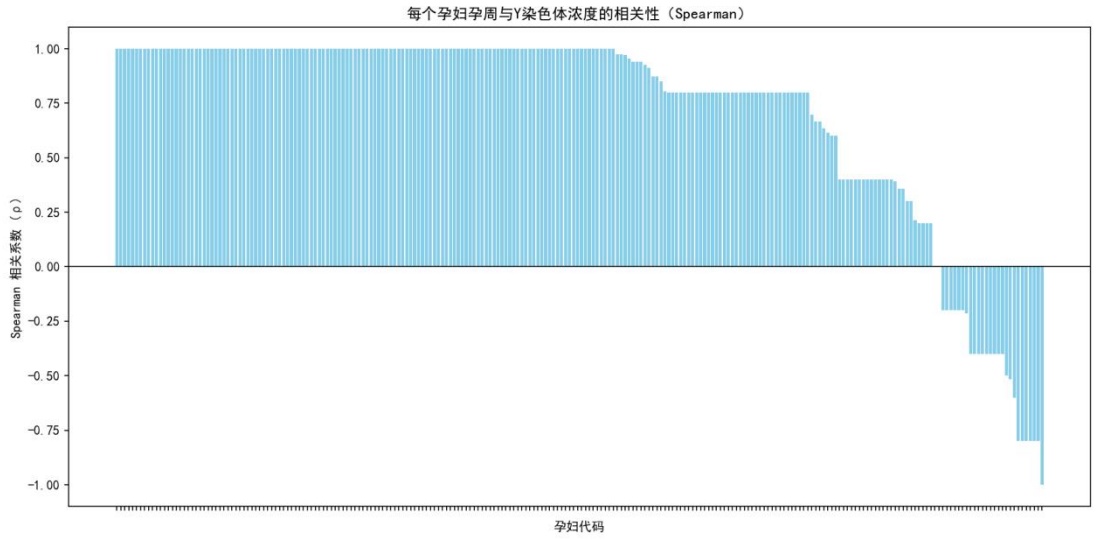


图 6：各孕妇孕周与 Y 染色体浓度的相关性

基于相关性分析结果和临床意义，我们最终选择了 3 个核心预测变量：标准化后的孕周数（w_i^*）、分位数归一化后的 BMI（B_i^*）和年龄（A_i）。

5.1.3 GAM 模型选择的理由

通过对热力图分析，我们首先考虑到 Y 染色体浓度与孕周、BMI、年龄以及怀孕次数之间存在较强的关系。通过对散点图的探索，我们还注意到 Y 染色体浓度并不是随着孕周程线性增长关系，而且 BMI 对浓度的影响更为复杂，不同区间呈现出明显的差异。这些图像表明，简单的线性模型难以准确描述数据背后的规律。另一方面，年龄和怀孕次数虽然也会对 Y 染色体浓度影响产生影响，但更符合线性累加的特征。因此，如果我们把这些变量放在同一个框架里进行建模，就需要一种既能处理非线性关系，又能兼顾线性解释性的工具，为此我们选

择广义加性模型（GAM)。

GAM 模型允许我们用平滑函数去拟合孕周和 BMI 与浓度之间的非线性关系，同时又保留了年龄的线性项，从而在保证模型灵活性的同时维持了解释的直观性。另外，孕周与 BMI 之间还存在交互作用，高 BMI 孕妇在不同孕周的变化趋势与低 BMI 孕妇有显著差异，这种二维关系用 GAM 的张量积平滑就能够很好地捕捉。此外，Y 染色体浓度的数据分布并非正态。经 K-S 检验，Gamma 分布的 p 值为 0.34，表明可以认为其服从 Gamma 分布。GAM 天然支持指数族分布的建模，这让它在拟合时更贴合数据的真实特征。

5.1.4 GAM 模型的基本框架

对孕妇 i，我们记其检测得到的 Y 染色体的浓度为 $\mu_i = E[Y_i]$。由于 BMI、检测孕周和年龄对染色体浓度均呈非线性关系，所以我们分别用平滑函数$f_1(B_i)$、$f_2(w_i)$、$f_3(A_i)$来刻画他们对染色体的影响。对于孕周、BMI 和年龄的交互效应，我们采用张量积平滑函数$f_{12}(w_i^*, B_i^*)$、$f_{23}(B_i^*, A_i)$、$f_{13}(w_i^*, A_i)$作为衡量它对 Y 染色体的影响函数。因此，我们的 GAM 模型为：

$$[\log(\mu_i) = \beta_0 + f_1(w_i^*) + f_2(B_i^*) + f_3(A_i) + f_{12}(w_i^*, B_i^*) + f_{23}(B_i^*, A_i) + f_{13}(w_i^*, A_i)]$$

对于平滑函数，采用 B 样条基展开：

$$f_j(x) = \sum_{k=1}^{K_j} \gamma_{jk} B_{jk}(x)$$

对于交互项，采用张量积形式：

$$f_{12}(w^*, B^*) = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{12,k_1k_2} B_{1k_1}(w^*) B_{2k_2}(B^*)$$

其中$B_{jk}(\cdot)$为三次 B 样条基函数，γ_{jk}为待估参数。

5.1.5 GAM 模型的求解

5.1.5.1 估计参数

为了估计参数，采用惩罚似然估计方法，目标函数为：

$$L(\theta) = \ell(\theta) - \frac{1}{2}\sum_{j} \lambda_j \gamma_j^T \mathbf{S_j} \gamma_j$$

因为响应变量 Y 染色体浓度符合 Gamma 分布，因此我们将$\ell(\theta)$定为$Gamma$分布的对数似然函数，λ_j为平滑参数，$\mathbf{S_j}$为惩罚矩阵，控制平滑函数的粗糙度。

5.1.5.2 选择平滑参数

其次在平滑参数的选择上，我们采用限制极大似然（REML）方法自动选择最优平滑参数λ_j，目的是在拟合优度与模型复杂度之间寻求最优平衡。

5.1.5.3 迭代求解

使用 Fisher 评分算法进行迭代求解：

$$\beta^{(k+1)} = (\mathbf{X}^T \mathbf{W}^{(k)} \mathbf{X} + \mathbf{S}_\lambda)^{-1} \mathbf{X}^T \mathbf{W}^{(k)} \mathbf{z}^{(k)}$$

不断迭代至收敛准则$\frac{||\beta^{(k+1)} - \beta^{(k)}||}{||\beta^{(k)}||} < 10^{-6}$满足。

由于生物数据存在测量误差和个体误差，我们将置信区间设定为 95%，并将落在该区间内的预测值均视为预测正确。

5.1.5.4 结果分析

本题使用 python 代码实现，最终得到结果：

选择的分布	Gamma
拟合成功率	60.05%
Pseudo R^2:	0.5123

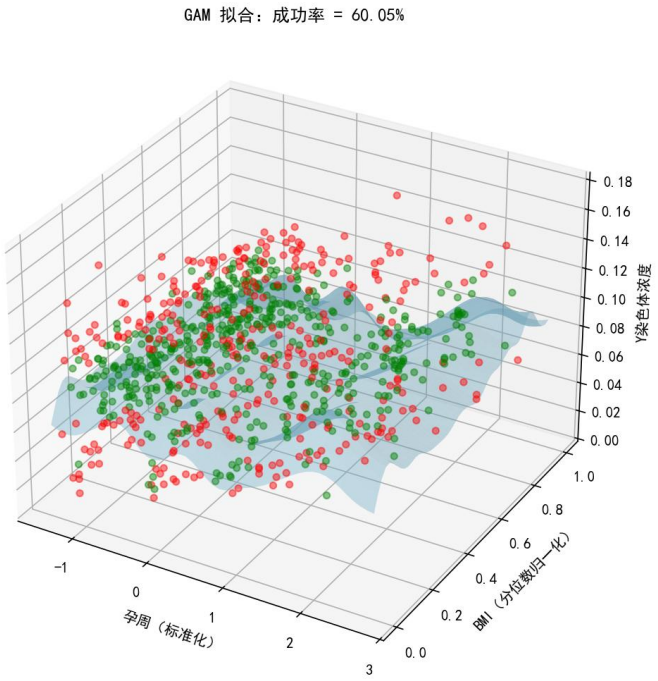


图 7：三维散点拟合图

该模型成功率为 60.05%，说明模型能够在大部分样本上实现有效拟合；Pseudo R^2 为 0.5123，意味着模型可以解释约一半的数据变异性。整体来看，模型能够较好地捕捉数据的主要趋势和模式，反映了孕周、BMI、年龄对 Y 染色体浓度的显著影响，为后续分析和预测提供了可靠依据。

5.1.6 模型显著性检验

为了验证 GAM 模型中各项对 Y 染色体浓度的影响是否显著，我们采用 F 检验进行显著性检验。p 值是 F 检验中衡量显著性水平的指标，p 值越小，代表模型的显著性越高，即该项对 Y 染色体浓度的影响越显著。对于 F 统计量计算公式为：

$$F_j = \frac{(\text{Deviance}_{\text{reduced}} - \text{Deviance}_{\text{full}})/\text{edf}_j}{\text{Deviance}_{\text{full}}/\text{df}_{\text{residual}}}$$

其中：

$Deviance_{reduced}$为去掉第 j 个平滑项后简化模型的偏差；

$Deviance_{full}$ 为包含所有平滑项的完整模型的偏差；

edf_j为第 j 个平滑项的有效自由度；

$df_{residual}$为完整模型的残差自由度。

根据统计学惯例，当 $p < 0.05$ 时，我们可以说明这个模型是显著的。通过对 GAM 模型各项进行 F 检验，得到 F 值和 p 值分布表。得到如下显著性检验结果：

变量	系数	标准误(SE)	z 值	P 值
孕周(标准化)	-0.1211	0.2175	-0.5569	1.214e-09
BMI(分位数归一化)	-0.2574	0.0839	-3.0683	0.0002257
年龄	-0.03719	0.0073	-4.8138	9.007e-05

5.2 问题二：BMI 分组与最佳 NIPT 时点优化

5.2.1 模型的建立

将孕妇按 BMI 进行合理分组、确定每组最佳检测时点、最小化整体潜在风险，这是一个涉及分组数量、分组边界和检测时点的多变量组合优化问题。同时我们需要考虑假阴性风险、时间窗口风险等多个因素对检测准确性的影响。又因为 BMI 分组必须是连续区间且不重叠，每组内的孕妇采用统一的检测时点，这使得问题具有明显的组合优化特征。

5.2.1.1 定义决策变量

假设我们有 N 个孕妇，每个孕妇 i 有 BMI 值B_i、年龄A_i等属性。我们要将这些孕妇按 BMI 分成 G 组，并为每组确定最佳检测时点。

（1）分组决策变量：$Z_{i,g} \in \{0,1\}$，表示孕妇 i 是否被分配到第 g 组，

(2) 检测时点决策变量：$U_{g,w} \in \{0,1\}$，表示第 g 组的检测时点是否为第 w 周

(3) 个体检测决策变量：$S_{i,w} \in \{0,1\}$，表示孕妇 i 是否在第 w 周进行检测

(4) 组数决策变量：$G \in \{2,3,4,5,6\}$，表示总共分为 G 组

(5) BMI 分界点决策变量：$c_j \in [26,47], j = 1,2,\ldots,G-1$，约定$c_0 = 26$，$c_G = 47$

(6) Y 染色体浓度预测值辅助变量：$\hat{Y}_{i,w*}$

5.2.1.2 变量关系约束

(1) 分组唯一性约束：每个孕妇必须且只能被分配到一个组

$$\sum_{g=1}^{G} Z_{i,g} = 1, \ \forall i = 1,2,\ldots,N$$

(2) 检测时点唯一性约束：每个组必须且只能有一个检测时点

$$\sum_{w=14}^{29} U_{g,w} = 1, \ \forall g = 1,2,\ldots,G$$

(3) 分组逻辑约束：根据 BMI 值确定分组归属

$$Z_{i,g} = 1 \Leftrightarrow c_{g-1} \le b_i < c_g, \ \forall i = 1,2,\ldots,N, g = 1,2,\ldots,G$$

(4) 检测决策关联约束：个体检测决策由其分组和组检测时点共同决定

$$S_{i,w} \ge Z_{i,g} + U_{g,w}, \ \forall i = 1,2,\ldots,N, w = 12,\ldots,32, g = 1,2,\ldots,G$$

$$S_{i,w} \le Z_{i,g}, \ \forall i = 1,2,\ldots,N, w = 12,\ldots,32$$

$$S_{i,w} \le U_{g,w}, \ \forall i = 1,2,\ldots,N, g = 1,2,\ldots,G$$

(5) 组内 BMI 范围约束：每组 BMI 的范围不能小于 3

$$c_j - c_{j-1} \ge 3, j = 1,2,\ldots,G$$

5.2.1.3 假阴性风险函数

基于问题一建立的 GAM 模型，我们可以预测孕妇 i 在孕周 w 的 Y 染色体浓度$\hat{Y}_{i,w}$。由题意 Y 染色体浓度小于 4%，则可以判断该检测不具有准确性。因此采用分段函数来描述 Y 染色体浓度与假阴性概率的关系：

$$FN_i = \begin{cases} 1, & \hat{Y}_{i,w^*} < 0.04 \\ \exp\left(-50(\hat{Y}_{i,w^*} - 0.04)\right), & \hat{Y}_{i,w^*} \geq 0.04 \end{cases}$$

该模型采用指数衰减函数是因为当 Y 染色体浓度略高于 4%阈值时，仍存在一定的假阴性风险，但这种风险随着浓度的增加而快速降低。参数 50 的选择使得浓度达到 5%时假阴性概率降至约 0.37，符合临床观察。

5.2.1.4 时间窗口风险函数

时间窗口风险的建模基于问题描述中的风险等级划分。早期发现（12 周以内）、中期发现（13-27 周）和晚期发现（28 周以后）分别对应不同的基础风险系数，且在每个阶段风险均会随时间增长而小幅度抬升：

$$R_{i,w^*} = \begin{cases} 1 + 0.05 \times w^*, & \text{if } w^* \leq 12 \\ 2 + 0.02 \times (w^* - 13), & \text{if } 13 \leq w^* \leq 27 \\ 3 + 0.01 \times (w^* - 28), & \text{if } w^* \geq 28 \end{cases}$$

5.2.1.5 构建目标函数和约束条件

为了综合考虑假阴性风险、时间窗口风险、分组复杂度和区间合理性，我们构建如下目标函数：

$$\min F = \lambda_1 \sum_{i=1}^{N} FN_i + \lambda_2 \sum_{i=1}^{N} R_{i,w^*}$$

其中，第一项$\sum_{i=1}^{N} FN_i$衡量总体假阴性风险，权重λ_1确保检测准确性的重要性；第二项$\sum_{i=1}^{N} R_{i,w*}$衡量时间窗口风险，权重λ_2平衡早期检测的需求。在对权重参数的设置上，检测准确性是首要考虑因素，时间风险次之，分组简洁性和区间合理性则作为约束条件。通过多次试验调整，确保不同目标之间的平衡。

$$
\text{s.t.}\begin{cases}
\sum_{g=1}^{G} Z_{i,g} = 1, \forall i = 1,2,\ldots,N \text{(每个孕妇必须且只能被分配到一个}BMI\text{组)} \\
\sum_{w=14}^{29} U_{g,w} = 1, \forall g = 1,2,\ldots,G \text{(每组只能有一个检测时点)} \\
Z_{i,g} = 1 \Leftrightarrow c_{g-1} \le b_i < c_g,\ \forall i = 1,2,\ldots,N, g = 1,2,\ldots,G \text{(BMI 分组逻辑)} \\
\left.\begin{aligned}
S_{i,w} \ge Z_{i,g} + U_{g,w},\ \forall i = 1,2,\ldots,N, w = 12,\ldots,32, g = 1,2,\ldots,G \\
S_{i,w} \le Z_{i,g},\ \forall i = 1,2,\ldots,N, w = 12,\ldots,32 \\
S_{i,w} \le U_{g,w},\ \forall i = 1,2,\ldots,N, g = 1,2,\ldots,G
\end{aligned}\right\} \text{(检测决策逻辑关联)} \\
c_j - c_{j-1} \ge 3, j = 1,2,\ldots,G \text{(每个 BMI 组最小组间距保证区分度)} \\
26 = c_0 < c_1 < c_2 < \ldots < c_{G-1} < c_G = 47 \text{(分界点具有单调性)} \\
w \in [12,32] \text{(检测周在可行范围内)} \\
G \in \{2,3,4,5,6\} \text{(分组数量适中)} \\
Z_{i,g}, U_{g,w}, S_{i,w} \in \{0,1\}, c_j \in R^+ \text{(变量类型)}
\end{cases}
$$

5.2.2 遗传算法求解

本问题属于混合整数线性规划，包含二元变量 $Z_{i,g}$，$U_{g,w}$，$S_{i,w}$、连续变量 c_j 以及整数变量 G，且约束条件复杂。传统数学规划方法难以直接求解，因此采用遗传算法进行求解。

5.2.2.1 适应度函数

适应度函数基于目标函数设计，同时将部分约束条件作为惩罚加入。根据假阴性风险和时间窗风险的相对尺度和优先程度将原目标函数中两项风险的系数确定为$\lambda_1 = 10$，$\lambda_2 = 0.4$。适应度函数的整体设计及计算过程如下：

Step 1: 根据(cuts, weeks)确定每个孕妇的分组和检测时点

$$
\text{group}_i = \max\{g : b_i \ge c_{g-1}\},\ w_i^* = w_{\text{group}_i}
$$

Step 2:基于 GAM 模型预测 Y 染色体浓度 $\hat{Y}_{i,w_i^*} = GAM(g_{i,w_i^*}^{std}, b_i^{norm}, a_i)$

Step 3: 计算假阴性风险 $FN_i =$

$$\begin{cases} 1 & if \hat{Y}_{i,w_i^*} \leq 0.04 \\ \exp(-50 \times (\hat{Y}_{i,w_i^*} - 0.04)) & if \hat{Y}_{i,w_i^*} > 0.04 \end{cases}$$

Step 4: 计算时间窗口风险 $R_i = \begin{cases} 1 & if w_i^* \leq 12 \\ 1.5 & if 12 < w_i^* \leq 27 \\ 5 & if w_i^* > 27 \end{cases}$

Step 5: 计算区间惩罚 $P_j = \begin{cases} 100 & if c_j - c_{j-1} < 3 \\ 0 & otherwise \end{cases}$

Step 6：计算得到最终适应度：$f(cuts, weeks) = 10 \sum_{i=1}^{N} FN_i + 0.4 \sum_{i=1}^{N} R_i + P_j$

5.2.2.2 染色体编码

本算法采用双基因变长实数编码。每个个体由两端基因组成，第一段基因是 BMI 组的分界点，长度随组数而变化；第二段是检测孕周，与组数长度一致，表示每组的检测孕周。这种编码方式既能灵活表示不同组数的分组方案，又能直接操作实数值，方便遗传算法进行变异和交叉操作。

5.2.2.3 种群初始化

Step1：随机生成分组数：$G \in \{2,3,4,5,6\}$

Step2：在约束条件下随机生成切点向量：$cuts \sim U[26, 47 - 3(G-1)]$，排序后加间距修正 ，保证每组的 BMI 区间不小于 3。

Step3：随机生成检测周向量：$weeks \sim U\{12, 13, \ldots, 32\}$

5.2.2.4 选择操作

选择操作能在当前种群中挑选优秀个体作为父代，将该代的优秀基因传递给下一代。常见的选择方法包括轮盘赌选择、锦标赛选择和排名选择等。该算法采用基于适应度的轮盘赌选择，具体公式为：

$$P(select_k) = \frac{\exp(-f_k)}{\sum_{j=1}^{POP\_SIZE} \exp(-f_j)}$$

其中f_k为第 k 个个体的适应度值。

5.2.2.5 变异操作

切点变异：以概率p_{mut}对每个切点进行高斯扰动 $c_j^{new} = c_j^{old} + \mathcal{N}(0,1)$ 变异后重新排序并修正间距约束

检测周变异：以概率p_{mut}随机重选检测周 $w_g^{new} \sim U\{12,13,\ldots,32\}$

5.2.2.6 交叉操作

在遗传算法中，交叉操作用于生成新的个体。本算法选择了双基因变长交叉，先将父代的 BMI 分界点合并并去重，形成一个分割点集合。然后随机确定子代的组数，并从合并后的分割点中选择对应数量的 cut 作为子代的 BMI 分界点。同时，为子代随机分配检测孕周，确保其数量与组数匹配。

5.2.3 结果分析

本小问将孕妇数据带入上述遗传算法模型，求解得出应按照 BMI 分为 4 组。具体的 BMI 分组情况及最佳 NIPT 时点如下：

BMI 分组区间	[26.0,34.5)	[34.5-38.2)	[38.2,41.2)	[41.2,47)
最佳 NIPT 时点	14	14	16	17

散点图的纵坐标是原样本孕妇 Y 染色体浓度首次大于 4%的孕周，散点颜色代表了 BMI 分组，红色虚线表示了每一组的推荐检测孕周。从中可清晰观察到，大部分孕妇在最佳 NIPT 时点前 Y 染色体浓度能够大于 4%。

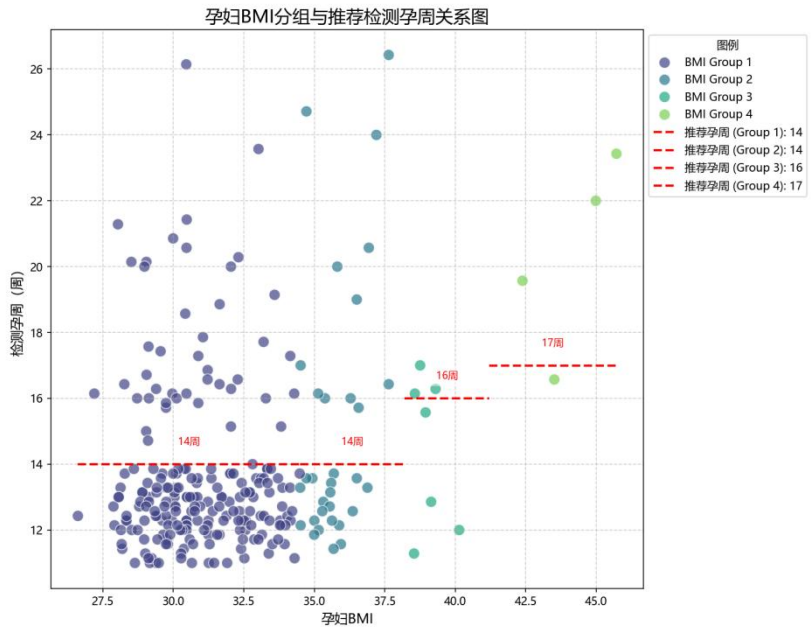


图 8：孕妇 BMI 分组与推荐检测孕周关系图

尽管大部分孕妇在推荐的 NIPT 检测时点前可检测到 Y 染色体浓度超过 4%，但仍有少部分孕妇在最佳检测时点之后才能检测到该浓度。考虑到检测时间窗可能带来的风险，从医学角度仍建议尽早进行检测。

在第 4 组中，仅有一名孕妇可在推荐的 NIPT 检测时点前达到该浓度，其余未达到推荐标准。这三名孕妇代码分别是 A099，A118，A163。A118 孕妇仅在 22 周检测了一次；A099 和 A163 孕妇在 17 周前最近一次检测的 Y 染色体浓度分别为 0.37 和 0.34，均接近 4%。从医学角度来看，相比多次重复检测的成本，尽早进行检测降低检测时间窗风险更重要。因此该最佳 NIPT 检测时点是合理的。

5.2.4 检测误差对模型的影响

检测结果可能受到人员操作、检测仪器、检测环境等多种因素影响，这会导致检测所得的 Y 染色体浓度不准确，进而影响判断 NIPT 的结果是否准确。为了模拟检测误差对模型造成的影响，我们在通过 GAM 模型预测得到的 Y 染色体浓度上增加不同强度的高斯噪声。

噪声分布	BMI 分组结果	每组最佳 NIPT 检测时点
无噪声	[26.0,34.5)，[34.5-38.2)，[38.2,41.2)，[41.2,47.0)	14,14,16,17
$N(0,(1\times10^{-3})^2)$	[26.0,34.5)，[34.5-38.2)，[38.2,41.2)，[41.2,47.0)	14,15,15,17
$N(0,(2\times10^{-3})^2)$	[26.0,34.5)，[34.5-38.3)，[38.3,41.9)，[41.9,47.0)	14,16,22,25

当添加的高斯噪声标准差为 0.001 时，Y 染色体浓度检测误差的浮动范围约在 ± 0.3%，该误差对 BMI 分组不产生影响，每组最佳 NIPT 检测时点有小幅度的变化。当噪声标准差为 0.002 时，Y 染色体浓度检测误差的浮动范围约在 ± 0.6%，该误差对 BMI 分组几乎不产生影响，但每组最佳 NIPT 检测时点变化幅度较大。通过查阅文献，Y 染色体浓度的检测误差一般在 0.2%到 0.5%之间，在考虑了真实范围内的检测误差后，模型的结果没有并发生重大改变，这也说明了该模型的鲁棒性较强。

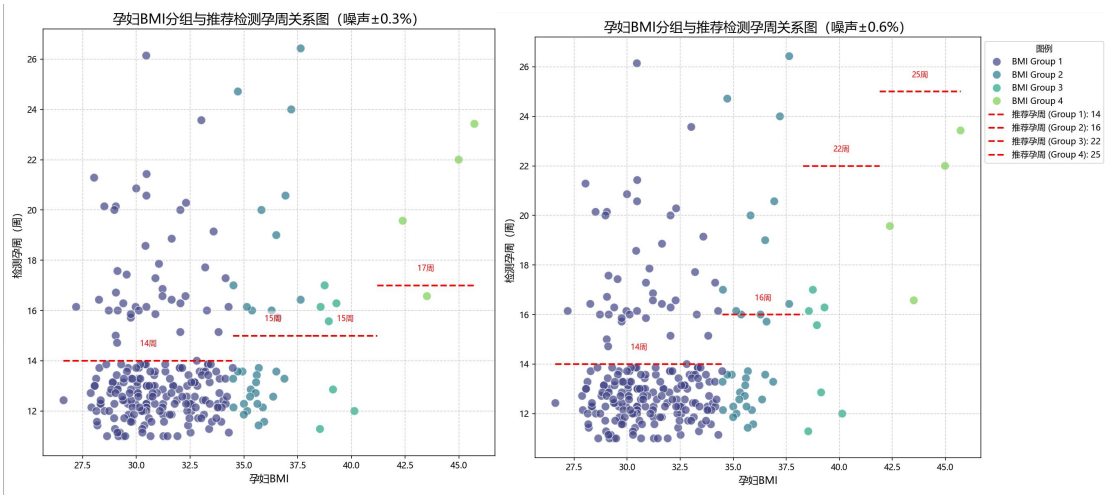


图 9：孕妇 BMI 分组与推荐检测孕周关系图（加噪声）

5.3 问题三：多因素综合优化模型

5.3.1 模型的建立

在问题二的基础上，本问题旨在构建一个更全面的优化模型，以确定男胎孕妇的 BMI 分组和 NIPT 时点。除了 BMI 和孕周这两个核心变量，我们还纳入了其他可能影响 Y 染色体浓度的因素，包括 IVF 妊娠类型、怀孕次数、生产次数、年龄和身高。这些变量的引入，能更准确地反映临床实践中的复杂情况，为检测方案提供更加完善的决策支持。

我们依然采用 GAM 模型作为预测核心，但对模型结构和变量进行了改进。

5.3.1.1 增强型 GAM 模型构建

Step1：变量筛选与多重共线性处理

在变量选择过程中，我们首先分析了体重与 BMI 之间的相关性。由于 BMI = 体重/身高²，，通过数据分析，我们得到的皮尔逊相关系数显示体重与 BMI 的相关系数为 0.84，存在严重的多重共线性问题。又考虑到第一问 GAM 模型分析中 BMI 平滑项具有显著性，因此我们选择保留 BMI 而排除体重因素在模型之外，以免影响模型准确性。

Step2：类别变量处理

在本问中视 IVF 妊娠类型、怀孕次数、生产次数为类别变量，并采用独热编码方式将其转化成数值形式：

$$IVF_{ij} = \begin{cases} 1, & \text{if } i\text{的 IVF 类型为} j \\ 0, & \text{otherwise} \end{cases}$$

类似地定义P_{ik}和$Birth_{il}$。 这些独热编码后的变量将会作为线性项被纳入 GAM 模型中，以更准确地部抓他们对 Y 染色体浓度的影响。

Step3：扩展 GAM 模型架构

在问题一 GAM 模型的基础上，我们纳入额外的影响因素，构建增强型 GAM 模型：

$$\log(\mu_i) = \beta_0 + f_1(w_i^*) + f_2(B_i^*) + f_{12}(w_i^*, B_i^*) + \beta_3 A_i + \beta_4 H_i + \beta_{cat}^T X_{cat,i}$$

其中：H_i为孕妇i的身高；$\beta_{cat}^T X_{cat,i}$为由 IVF 妊娠类型、怀孕次数和生产次数经过独热编码后得到的类别变量向量。

5.3.1.2 Y 染色体达标率建模

Step1：达标阈值定义

由题意得当 Y 染色体浓度≥4%时认为达标，为此，我们首先使用 GAM 模型计算出在不同 BMI 分组和检测时点下，每个孕妇的 Y 染色体浓度预测值。接着，我们基于这些预测值，计算出每个分组的 Y 染色体浓度达标比例。因此对于孕妇i在孕周w的达标概率为：

$$P_{\text{达标}}(i,w) = P(\widehat{Y}_{i,w} \geq 0.04)$$

Step2：群体达标率计算

对于特定孕周 w，群体达标率定义为：

$$R_{\text{达标}}(w) = \frac{1}{N}\sum_{i=1}^{N} I(\widehat{Y}_{i,w} \geq 0.04)$$

其中$I(\cdot)$为指示函数。

Step3：达标率风险调整

将群体达标率纳入风险评估，构建达标率调整因子：

$$R_{\text{达标调整}}(w) = \begin{cases} 0.5, & if\ R_{\text{达标}}(w) < 0.7 \\ 0.2, & if\ 0.7 \leq R_{\text{达标}}(w) < 0.9 \\ 0, & if\ R_{\text{达标}}(w) \geq 0.9 \end{cases}$$

5.3.1.3 GC 含量质量评分机制

Step1：GC 含量质量评分函数

由题意，正常 GC 含量应维持在 40%-60%范围内，GC 含量的显著偏离通常指示测序过程中存在技术偏差或样本质量问题。通过对本研究数据集的分布特征分析，发现样本 GC 含量主要集中在 39%附近，略低于标准范围下限。

考虑到实际检测环境中的技术限制和样本特性，我们构建了基于 GC 含量的测序质量评价体系。该体系采用分层奖励机制，具体如下：首先，对于 GC 含量位于标准范围内（40%-60%）的样本，赋予最高质量权重；其次，考虑到本数据集的分布特征和临床检测的实际情况，对 GC 含量在 38%-42%范围内的样本也给予质量认可，但权重相对较低。

根据测序质量标准，建立 GC 含量评分机制：

$$GC_{score}(gc_i) = \begin{cases} 1, & if\ 38 \le gc_i \le 42 \\ 0.5, & if\ 40 \le gc_i \le 60 \\ 0, & otherwise \end{cases}$$

Step2：质量调整因子

群体平均 GC 质量得分为：

$$\overline{GC_{score}} = \frac{1}{N}\sum_{i=1}^{N} GC_{score}(gc_i)$$

质量惩罚项定义为：

$$Penalty_{GC} = 2 \times (1 - \overline{GC_{score}})$$

5.3.1.4 增强目标函数构建

Step1：改进的假阴性风险函数

结合噪声影响，假阴性风险函数修正为：

$$FN_i = \begin{cases} 1, & if\ \widehat{Y}_{i,w_i^*} + \epsilon_i \le 0.04 \\ exp\left(-50\left(\widehat{Y}_{i,w_i^*} + \epsilon_i - 0.04\right)\right), & if\ \widehat{Y}_{i,w_i^*} + \epsilon_i > 0.04 \end{cases}$$

其中$\epsilon_i \sim N(0,0.0005^2)$为检测误差。

Step2：改进的时间窗口风险函数

考虑 BMI 对时间风险的影响，修正时间风险函数：

$$R_{i,w^*} = \begin{cases} 1 + 0.05w^*, & w^* \le 12 \\ 2 + 0.02(w^* - 13), & 13 \le w^* \le 27 \\ 3 + 0.01(w^* - 28), & w^* \ge 28 \end{cases}$$

对于 BMI > 35 的孕妇，增加额外惩罚：

$$R_{i,w^*}^{adj} = R_{i,w^*} + 0.1 \times (B_i - 35) \times I(B_i > 35)$$

Step3：综合目标函数

该目标函数的建立不仅基于增强型 GAM 模型对 Y 染色体浓度的预测结果，还进一步结合了包含噪声项和检测误差的假阴性风险、BMI 调整后的时间窗口风险、达标率修正因子以及 GC 含量质量惩罚项，从而使得适应度能够更全面地反映检测方案的优劣。我们构建包含多因素考量的综合目标函数如下：

$$\min F = 10\sum_{i=1}^{N} FN_i + 2\sum_{i=1}^{N} R_{i,w^*}^{adj} + \sum_{w=14}^{29} R_{\text{达标调整}}(w) \times \sum_{i:w_i^*=w} 1$$

$$+ Penalty_{GC} - 0.5G + \sum_{g=1}^{G-1} P_g$$

5.3.2 改进遗传算法求解

Step1：适应度函数更新

采用优化后的目标函数作为适应度函数，更新适应度计算流程，根据增强 GAM 模型预测 Y 染色体浓度：

$$\hat{Y}_{i,w_i^*} = exp(\beta_0 + f_1(w_i^*) + f_2(B_i^*) + f_{12}(w_i^*, B_i^*) + \beta_3 A_i + \beta_4 H_i +$$

类别项)

1. 计算假阴性风险FN_i（含噪声）
2. 计算 BMI 调整的时间风险$R_{i,w*}^{adj}$
3. 计算达标率调整因子$R_{达标调整}(w)$
4. 计算 GC 质量惩罚$Penalty_{GC}$

Step2：遗传操作改进

保持原有遗传算法框架，但在适应度评估中使用新的目标函数。变异操作中增加对 GC 质量分布的考虑：

变异概率自适应调整：

$$p_{mut}^{new} = p_{mut} \times (1 + 0.1 \times (1 - \overline{GC_{score}}))$$

Step3：约束条件更新

在原有约束基础上，增加质量约束：

$$\overline{GC_{score}} \geq 0.6$$

5.3.3 结果分析

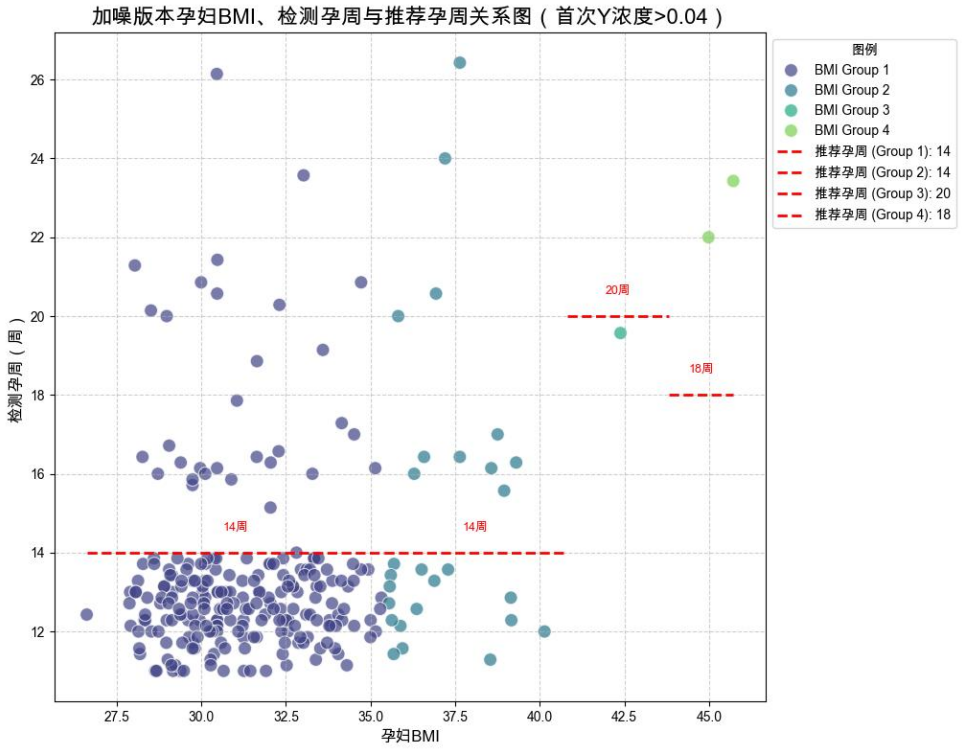


图 10：加噪版孕妇 BMI、检测孕周与推荐检测孕周关系图

给图为第三问改进后的 GAM+遗传算法模型输出的结果。横坐标是孕妇 BMI，纵坐标为检测孕周数。图中的散点代表每一位孕妇的 Y 染色体检测浓度首次达到 0.04 或以上的检测孕周，红线代表模型的 BMI 区间分组和推荐检测孕周结果。

在第一张图上可以看到，对于第一个区间[26,39.5]，模型的输出结果为 14 周进行 NIPT 检查最好。在这条线下分布着绝大多数检测点记录，意味着实际上大部分孕妇会在 14 周前做检测。这样的分布可以充分说明此模型对此 BMI 区间上的建议具有非常好的实际意义。对于区间[39.5,43.7]，模型推荐的时间（15 周）处于该类的 2 个数据点中间，说明模型学习到了相应的数据并综合考虑了样本点，即使数据点较少也能得出合理结果。对于最后一个区间[43.7,47]，模型的结果较检测记录早，其一可能是模型认为高 BMI 孕妇群体的 NIPT 时点过晚所带

来的潜在风险比提早检测的误检风险大，其二也可能是数据样本过少、分布较偏所致。总体来说，模型的结果较好，符合预期以及实践医学常理。

为了检测模型的泛化能力以及抗检测误差能力，我们还对模型进行了鲁棒性检验：对模型中最核心的中间变量 ypred 添加高斯噪声，并使用添加噪声后的数据输入遗传算法模型进行种群演化和结果预测，最终得到结果如下：

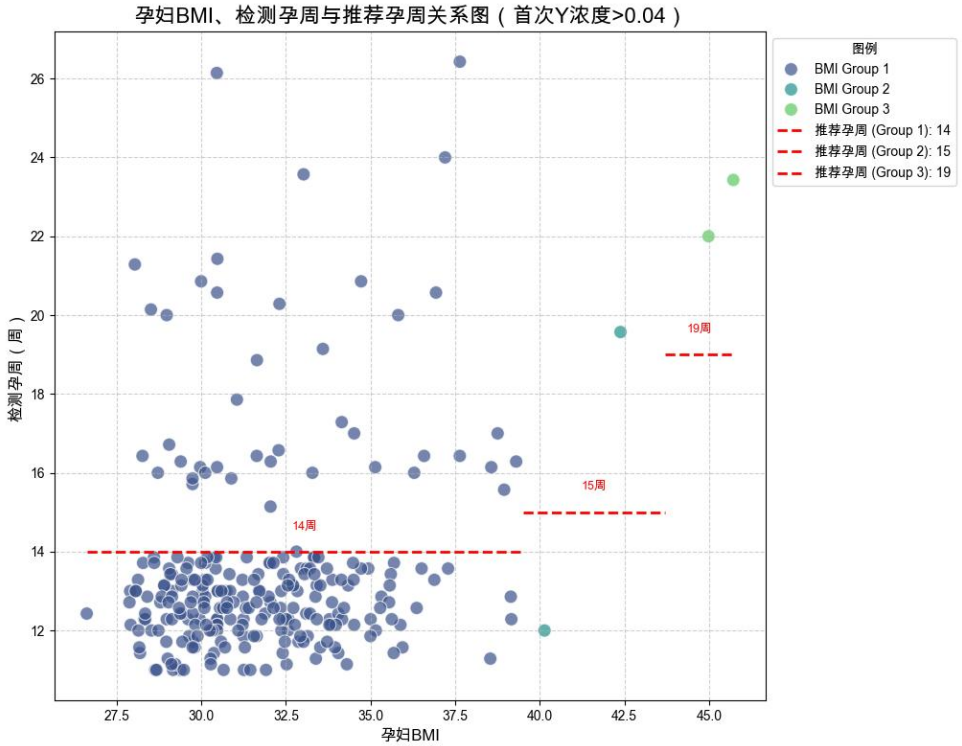


图 11：加噪版孕妇 BMI、检测孕周与推荐检测孕周关系图

可以看到，这次模型虽然多分出了一个区间，但是建议的检测周数与原本的分类结果相同，均为 14 周。这说明模型的主要部分并没有受到噪声影响。后半部分[40.8,43.8]区间结果为 20 周，可能为数据点样本被移动至前一区间，本区间样本过少导致的结果偏移；[43.8,47]区间结果为 18 周，与原本分类结果持平。综上所述，模型对 y 染色体浓度的检测误差具有较好的抗干扰性，在考虑检测误差的前提下可以使得结果相对稳定。

5.4 问题四：逻辑回归求解是否含有非整倍体分类问题

5.4.1 数据预处理

BMI 缺失值处理：存在一个 BMI 存在空缺的孕妇样本，使用该孕妇的身高

和体重计算 BMI 并进行填补。

分类变量数值化：将 IVF 妊娠情况和生产次数转化为数值型变量。

目标变量处理：将染色体非整倍体列中有数据的样本标记为“含非整倍体”，空值的样本标记为“整倍体”，并统一转化为数值型。

5.4.2 模型的建立

逻辑回归是一种解决分类问题的统计学方法，尤其适合解决二分类问题。在本问题中，我们选择逻辑回归来对染色体是否含有非整倍体进行分类。

逻辑回归模型会先将输入的变量按一定权重进行线性组合，再将线性组合的结果 z 输入到 Sigmoid 函数中，将其映射为 0 到 1 的概率，该概率表示了样本属于整倍体和非整倍体的可能性。

$$p = \sigma(z) = \frac{1}{1 + e^{-z}}$$

然后使用交叉熵损失函数衡量模型预测概率和真实类别的差距，用梯度下降法调整权重使损失函数最小，从而得到最优的模型。

5.4.3 模型的求解

为确定模型的输入变量，我们使用点二列相关系数衡量各个连续特征和染色体非整倍体之间的相关性，用 Cramer's V 量化分类特征和染色体非整倍体的关系强度。得到的结果如下：

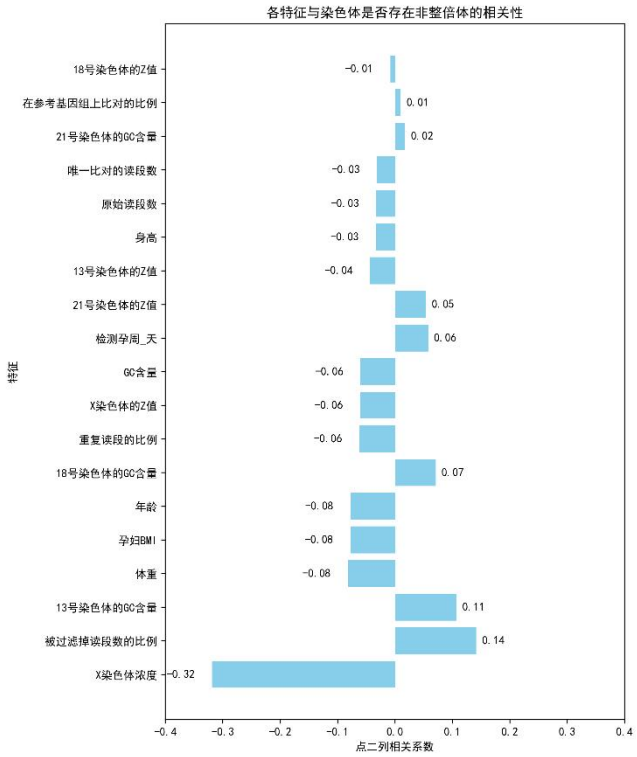
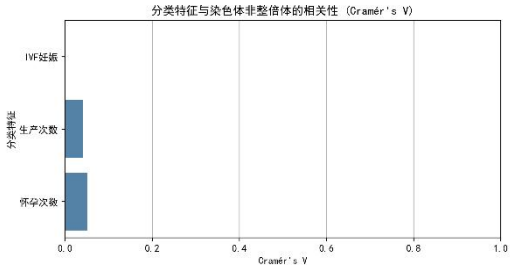


图 12：各特征与染色体是否存在非整数倍的相关性

结果显示，除了 X 染色体浓度和染色体非整倍体之间有弱相关性，其他特征的相关性均不明显，且三个分类变量的相关性极低，Cramer's V 均在 0.05 以下。基于此，我们将除了三个分类特征外的其他特征作为变量输入模型，以避免忽视这些特征对目标变量的影响。

通过统计发现，该数据集有非整倍体样本 67 个，整倍体样本 538 个。由于该数据集明显不平衡，直接对原数据进行逻辑回归可能导致对非整倍体样本的识



别能力低下。因此在逻辑回归前，我们先使用 SMOTETomek 方法进行过采样。SMOTETomek 可分为两步，首先使用 SMOTE 技术，在原有的非整倍体样本附近合成新样本，然后用 Tomek Links 去除噪声和模糊的边界样本。

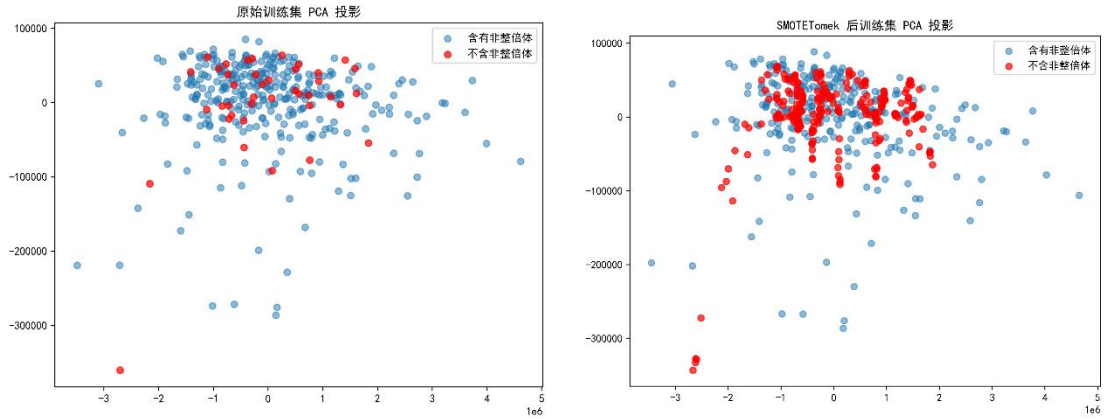


图 13：SMOTETomek 过采样前后 PCA 投影对比图

由于我们输入的变量较多，模型存在过拟合的风险，正则化的方式和强度会直接影响逻辑回归的效果。我们通过网格化搜索的方式评估不同正则化类型和强度在交叉验证集下的效果。得到的最优参数配置是 L1 正则化，正则化强度为 100。确定参数后，使用逻辑回归进行分类。

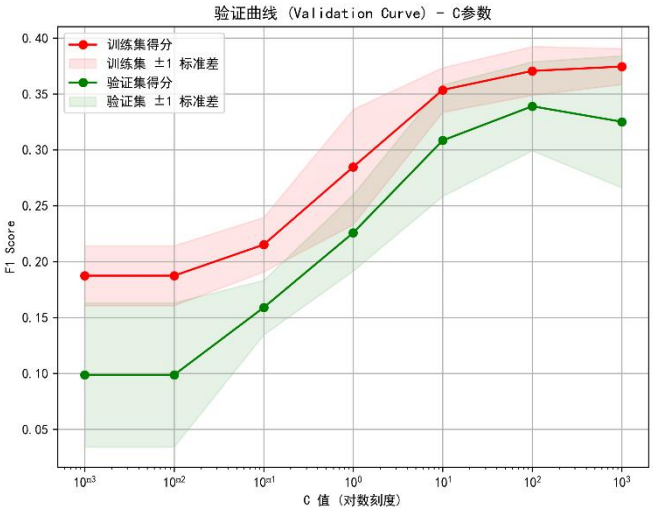


图 14：不同正则化强度下的 FI score 曲线

5.4.3 结果分析

这里使用 python 代码求解，在测试集上得到的结果如下：

样本	准确率	召回率	F1-score	样本数量
整倍体	0.97	0.82	0.89	215
非整倍体	0.36	0.78	0.49	27

该逻辑回归模型的 AUC 达到 0.86，整体分类能力较强。模型对整倍体的准确率和召回率都很高，虽然在非整倍体样本上的准确率相对较低，但召回率高达 78%，说明模型能够识别大部分非整倍体病例。在实际检测中，将有病状态误判为健康可能对孕妇造成严重危害，因此在保证召回率的前提下牺牲部分准确率是可以接受的策略。

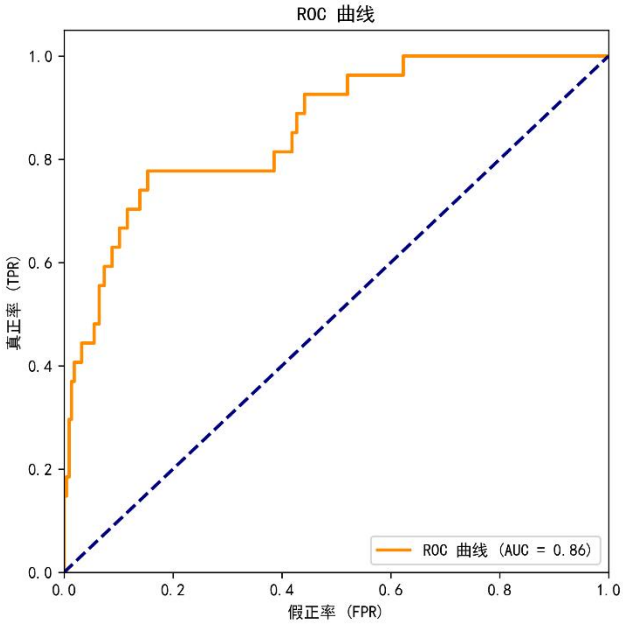


图 15：逻辑回归模型的 ROC 曲线