基于聚类分析的 NIPT 时点选择与胎儿的异常判定决策模型

摘要

无创产前检测（NIPT）作为现代产前筛查的重要技术，通过分析母体血液中胎儿游离 DNA 片段来检测染色体异常，为早期发现胎儿健康状况提供了有效手段。研究表明，胎儿 Y 染色体浓度与孕妇孕周和 BMI 密切相关，直接影响检测的准确性和临床风险。本文基于多元回归分析和机器学习方法，建立了 NIPT 检测时机优化与染色体异常判定的综合模型，为个性化产前筛查提供科学依据。

针对问题一，基于孕妇及其胎儿的检测数据，探讨孕妇 BMI、孕期周数等因素对男胎 Y 染色体浓度的影响。通过数据处理与相关性分析，建立了线性回归模型、多项式回归模型及混合效应模型进行数据分析，并利用残差图等工具进行模型诊断。计算结果表明，BMI 和孕周与 Y 染色体浓度之间相关系数较小但存在显著相关关系。

针对问题二，基于临床风险最小化原则，将 BMI 分为 5 组并确定最佳检测时点：BMI<28 组在孕 11-12 周，BMI 28-32 组在孕 13-14 周，BMI 32-36 组在孕 15-16 周，BMI 36-40 组在孕 17-18 周，BMI>40 组在孕 19-20 周，整体检测成功率从 72.4% 提升至 89.7%。

针对问题三，综合考虑身高、体重、年龄等多因素影响，建立了逻辑回归模型预测检测成功率，采用交叉验证优化参数，高 BMI 组成功率提升最为显著（从 58.3% 提升至 82.6%）。

针对问题四，由于女胎无 Y 染色体，所以通过 13 号、18 号、21 号染色体非整倍体检测结果为判定依据，综合考虑 Z 值、GC 含量、读段数、过滤比例、BMI 等多维因素，构建女胎异常风险预测模型。以 AB 列是否报告 T13/T18/T21 作为“异常”标签（1 表示异常，0 表示正常）。基于 604 例有效女胎样本，构建随机森林与逻辑回归模型进行分类预测。结果表明：逻辑回归模型表现更优，交叉验证 AUC 为 0.699，F1 为 0.285；随机森林 F1 仅为 0.077，表现较差。特征重要性分析显示，13 号染色体 GC 含量、孕妇 BMI、21 号染色体 GC 含量等质量与生理因素重要性高于 Z 值，可以见得 AB 列异常更可能由技术偏差或母体因素引起，而非胎儿真实异常。

关键词：相关性分析　多元回归分析　机器学习　随机森林　NIPT　检测优化　染色体异常判定

一、 问题重述

1.1 问题背景

无创产前检测（Non-invasive Prenatal Test, NIPT）是一种通过采集孕妇外周血，富集并测序胎儿游离 DNA（cfDNA），进而分析胎儿染色体是否存在非整倍体异常（如 21 三体、18 三体、13 三体）的先进技术。其核心原理是利用胎儿 DNA 在母体血浆中的存在比例进行统计推断。[?]

1.2 问题一

本题要求基于提供的孕妇NIPT检测数据，分析胎儿Y染色体浓度与孕妇孕周、BMI等指标的相关特性，建立相应的关系模型，并检验其显著性。具体包括：

1. 数据清洗与预处理；
2. 探索性数据分析（EDA），揭示变量间关系；
3. 构建 Y 染色体浓度与孕周、BMI 的数学模型；
4. 进行参数显著性检验与模型整体显著性检验；
5. 给出科学结论，为后续问题（如最佳检测时点）提供支持。

1.3 问题二

NIPT（无创产前检测）的准确性对男胎而言取决于 Y 染色体浓度是否 ≥4%，而检测时点的选择直接影响胎儿异常发现的风险（早期 ≤12 周风险低，中期 13-27 周风险高，晚期 ≥28 周风险极高）。题目明确 “BMI 是影响 Y 染色体浓度最早达标时间的主要因素”，需解决以下核心问题：

1. 对男胎孕妇按 BMI 进行合理分组；
2. 为每组确定最佳 NIPT 检测时点，在保证检测准确性（Y 浓度 ≥4% 的孕妇比例足够高）的前提下，最小化检测孕周，从而降低潜在健康风险；
3. 分析检测误差对结果的影响。

1.4 问题三

问题 2 仅基于 BMI 单因素和经验分位数确定 NIPT 检测时点，未考虑身高、体重、年龄等其他关键因素对 Y 染色体浓度达标的联合影响，可能导致时点推荐不够精准。问题 3 需解决以下核心任务：

1. 构建多因素预测模型，量化孕周、BMI、年龄、身高、体重、GC 含量等因素对 Y 染色体浓度达标概率（≥4%）的影响；
2. 模拟不同 BMI 组的“达标概率 - 孕周”曲线，为每组推荐达到 90% 达标概率（更高准确性要求）的“最佳 NIPT 时点”；
3. 分析检测误差（如浓度测量偏差）对模型预测结果及最佳时点的影响，验证模型稳健性。

1.5 问题四

在 NIPT 中，女胎因不携带 Y 染色体，传统基于 Y 染色体的胎儿 DNA 浓度评估失效，增加了异常判定的复杂性。题目要求：

1. 由于女胎无 Y 染色体，异常列全为“是”，需另寻判定依据；
2. 以 21 号、18 号、13 号染色体非整倍体（AB 列）为判定结果；
3. 综合考虑 X 染色体及上述染色体的 Z 值、GC 含量、读段数、过滤比例、BMI 等因素；
4. 建立女胎异常的判定方法。

由于 AE 列（胎儿是否健康）在女胎中全为“是”，无法作为真实异常标签，因此本文以 AB 列是否报告非整倍体（如 T21、T18、T13）作为“异常”标签，构建分类模型，探索影响异常判定的关键因素

二、 问题分析

2.1 问题一的分析

本题为分析男胎孕妇胎儿 Y 染色体浓度与孕周、BMI 的定量关系，首先通过数据预处理环节保障数据可靠性，针对关键变量，清洗 GC 含量超出 40%-60% 正常范围或测序深度过低的异常值，同时验证 BMI 计算一致性；接着开展探索性分析，通过绘制 Y 染色体浓度、孕周和 BMI 的分布图及散点图初步观察变量间关系趋势，计算多变量相关性热力图全面揭示变量间关联模式，为模型选择奠定基础；随后采用多层次建模策略构建数学模型：先以基础线性回归模型 $Y \sim GW + BMI$ 刻画线性关系，再引入多项式项（如 GW^2）和交互项（如 $GW \times BMI$）构建扩展模型以捕捉潜在非线性效应，提升模型稳健性；最后通过模型评估与验证确保模型有效性，借助 t 检验（参数显著性，$p < 0.05$）、F 检验（模型整体显著性）、R^2 与调整 R^2（拟合优度）评估模型性能，通过残差诊断（Q-Q 图、异方差性、自相关性检验）验证模型假设是否满足，以此建立准确描述胎儿 Y 染色体浓度与孕周、BMI 定量关系的数学模型并验证其统计显著性。

2.2 问题二的分析

本题的核心目标是确定不同 BMI 孕妇群体的最佳 NIPT 检测时点，以最小化与胎儿异常发现时间相关的潜在健康风险，该风险呈明显的时间依赖性，早期发现（≤12 周）风险低，中期（13–27 周）风险高，晚期（≥28 周）风险极高，而风险最小化的前提是保证检测准确性，对男胎而言需满足 Y 染色体浓度 ≥4%，因此目标函数被定义为在确保 Y 染色体浓度达标概率足够高（即检测准确）的前提下，通过最小化检测孕周来兼顾检测的时效性与可靠性。其中，关键变量为 BMI，这与问题 1 中 “BMI 与 Y 染色体浓度显著负相关” 的结论及题目明确指出的 “BMI 是影响 Y 染色体浓度最早达标时间的主要因素” 一致，意味着高 BMI 孕妇的 Y 染色体浓度增长更缓慢、达标时间更晚，需推迟检测，低 BMI 孕妇则可更早检测，故按 BMI 分组并差异化设定检测时点成为降低整体风险的必然策略。分组过程中需兼顾数据驱动性与临床可操作性，拟采用分位数法或 K-means 聚类分析，以 “Y 染色体浓度 ≥4% 的最早孕周” 为目标变量自动划分 BMI 切点，最终形成 3-5 个样本量均衡的组别，避免统计偏差并便于医疗实践执行。对于每个 BMI 分组，“最佳 NIPT 时点” 被定义为该组内 Y 染色体浓度首次达到或超过 4% 的第 p 百分位数孕周（p 通常取 80% 95%），例如某组 90% 的孕妇在 14 周时 Y 染色体浓度已达标，则建议该组最佳检测时点为 14 周，以此平衡检测的准确性（高 p 值）与时效性（低 p 值）。最后，为确保模型建议在实际应用中的可靠性，还需评估检测误差（如测序失败、生物学变异、测量不确定性）对推荐时点稳健性的影响，计划通过敏感性分析（如给浓度数据添加 ±5% 的噪声）和稳健性检验（如剔除低质量样本）等方法量化不确定性，并为最终的最佳时点提供置信区间估计。

2.3 问题三的分析

针对问题三，其核心是在问题二仅考虑 BMI 单因素的基础上，进一步纳入身高、体重、年龄等多维协变量，构建更全面的优化模型，目标为在保证 Y 染色体浓度达标比例（如 ≥90%）的前提下最小化检测时点，从而实现综合风险的最小化；相较于问题二，其复杂性显著提升，不仅从单因素（BMI）分组扩展到多因素协同分析，优化目标也从单一的 “最早达标时间” 转变为 “达标比例” 这一概率性指标，本质是需权衡 “早检测”（时效性）和 “准检测”（准确性）的多目标优化问题。解决该问题的关键在于建立 “达标比例” 与孕周及多因素间的定量关系，核心思路是为每一特定群体（如按 BMI 划分的组）绘制 “达标比例-孕周” 曲线，通过逻辑回归、Cox 比例风险模型等统计方法预测不同孕周及多因素条件下浓度达标的概率，例如构建 $log(\frac{p}{1-p}) = \beta_0 + \beta_1 \cdot GW + \beta_2 \cdot BMI + \beta_3 \cdot Age + \ldots$ 形式的模型，进而求解使达标概率 p 首次超过设定阈值（如 90%）的最小孕周，作为该组的推荐检测时点。同时，必须考虑检测误差的影响，因测量误差可能导致假阴性（实际浓度达标但测量值未达标）而延误干预时机，为此需对测量误差建模（如假设其服从正态分布 $N(0, \sigma^2)$），计算概率化的达标条件 $P(Y \geq 4\%) = 1 - \Phi(\frac{4\% - \hat{Y}}{\sigma})$，

增强模型对误差的稳健性。最终的风险最小化需通过综合框架实现，可定义风险函数 $R = w_1 \cdot (T - T_0)^+ + w_2 \cdot (1 - P_{acc})$（其中 $(T - T_0)^+$ 为延误早期发现窗口的时间成本，$(1 - P_{acc})$ 为检测失败的风险，w_1 与 w_2 为反映决策者偏好的权重系数），通过求解该函数的最小值，为不同特征的孕妇群体制定兼具时效性和可靠性的个性化检测方案。

2.4 问题四的分析

本题针对女胎染色体异常判定分析中，女胎数据总量为 605 例，经特征完整性筛选后得到有效样本 604 例，其中 AB 列（检测系统报警结果）非空的报告异常样本共 67 例，占比约 11.1%，呈现出显著的类别不平衡特征。我们分析的过程面临多重核心挑战，一方面是标签可靠性问题，AB 列作为检测系统输出的“报警结果”，可能存在假阳性情况，影响标签准确性；另一方面是特征维度高，数据涵盖染色体 Z 值、GC 含量、读段数、孕妇 BMI 等多类指标，需合理筛选有效特征；还有就是类别不平衡问题，异常样本仅占 11.1%，易导致模型学习偏向多数正常样本，降低异常检出能力；最后就是 Z 值核心性验证问题，理论上染色体 Z 值应为判定异常的最重要特征，但需通过实证分析验证其实际作用。针对上面提到的情况，我们决定采用监督学习方法展开：以 AB 列为判定标签构建分类模型，通过随机森林与逻辑回归两种算法的对比分析，结合特征重要性评估识别影响女胎染色体异常的关键因素，最终通过全面的模型性能评估，为临床女胎染色体异常判定提供科学合理的建议。针对女胎染色体异常判定的核心问题，结合数据特征，我们采用“数据预处理-特征工程-多模型构建-综合评估”的递进式流程。首先通过特征工程实现数据降维与质量提升，解决高维度与标签可靠性问题；随后构建多类监督学习模型，针对性处理类别不平衡等挑战；最终通过多指标评估体系，筛选最优模型并验证关键特征作用，形成科学的异常判定方案。

三、 模型假设

1. Y 染色体浓度与孕周、BMI 的关系可用线性或低阶多项式近似。
2. 不同孕妇之间的观测相互独立（但同一孕妇的多次检测存在相关性，通过混合模型处理）。
3. 残差方差在不同预测值下保持稳定（允许轻微异方差）。
4. 假设附件提供的 NIPT 数据真实可靠，测序质量指标（GC 含量、读段数、比对比例等）符合临床检测标准，数据缺失和异常值已在预处理中得到合理处理。
5. 假设假设孕妇 BMI、孕周等生理指标在检测期间相对稳定，胎儿 DNA 在母血中的比例变化主要受孕周和 BMI 影响，不考虑其他突发性生理变化或疾病因素的干扰。
6. 假设 Y 染色体浓度达到 4% 为 NIPT 检测准确性的可靠阈值，女胎 X 染色体浓度无异常即为正常，检测误差服从正态分布且可通过统计方法进行量化分析。

7. 假设早期发现（≤12 周）、中期发现（13-27 周）和晚期发现（≥28 周）的风险等级划分合理，风险最小化目标可通过数学优化方法实现，不考虑个体特异性风险偏好差异。

四、 符号说明

表 1　符号说明详

符号	说明	单位
Y_{conc}	Y 染色体浓度	%
BMI	身体质量指数	kg/m^2
GA	孕周	周
β_i	回归系数	-
ε	误差项	-
$P(success)$	检测成功概率	-
Age	孕妇年龄	岁
Z_{13}	13 号染色体 Z 值	-
Z_{18}	18 号染色体 Z 值	-
Z_{21}	21 号染色体 Z 值	-
Z_X	X 染色体 Z 值	-
GC_{13}	13 号染色体 GC 含量	%
GC_{18}	18 号染色体 GC 含量	%
GC_{21}	21 号染色体 GC 含量	%
$P(abnormal)$	染色体异常概率	-
w_i	特征权重	-
b	偏置项	-
$H(D)$	信息熵	-
IG	信息增益	-
AUC	ROC 曲线下面积	-
μ	均值	-
σ	标准差	-

注：其他文章内使用但未在表内详细说明的符号将在使用时给出说明。

符号	说明
Y	Y 染色体浓度（%）
W	孕周数（周）
B	孕妇 BMI（kg/m²）
A	孕妇年龄（岁）
G	GC 含量（%）
R	原始读段数
Z_i	第 i 号染色体 Z 值
$\beta_0, \beta_1, \ldots$	回归模型参数
r	相关系数
p	显著性 p 值
R^2	模型解释方差比例

表 2　符号说明表

五、 模型建立与求解

5.1 问题一模型的建立与求解

5.1.1 数据预处理

读取附件中男胎检测数据表和女胎检测数据表的所有数据，可得 1687 个初始总样本并转换数据类型，将孕周“周 + 天”转换为浮点数，使用公式“体重/(身高/100)^2”重新计算 BMI，把结果与原数据中“孕妇 BMI”列进行验证，计算最大差异为 0.0022，差异较小的进行忽略。以“Y 染色体浓度非空且 > 0”作为筛选条件筛选出男胎样本，可得 1082 个样本。对于筛选后的样本进行数据清洗，保留 GC 含量在 [0.35, 0.65] 范围内的数据。为保证测序深度，保留原始读段数大于 3,000,000，被过滤掉读段数比例小于 0.1。为有合理检测窗口，保留孕周在 [8, 28] 区间内；为排除极端值，剔除 Y 染色体浓度大于 15 的数据。观察数据表后，发现 BMI、孕周、Y 染色体浓度等关键指标的缺失值大概占 1.1%，就直接剔除这些缺失值。把末次月经日期和检测日期相减跟孕周比对，差别大的就当无效数据也剔除。对于非关键指标的缺失值，则采用均值或中位数填充。清洗后有效样本数为 925 个。

5.1.2 相关性分析

完成数据处理后，通过计算观察到数据的变量分布，为便于观察将相关变量的分布以散点图形式可视化，由图??和图??，可得 Y 染色体浓度的均值约为 0.15%，标准差约为 0.05%，呈右偏分布，部分样本浓度较高。孕周的均值约为 17.2 ，范围在 10 - 25 周，分布较为均匀。BMI 的均值约为 30.5，多数集中在 28 - 36。

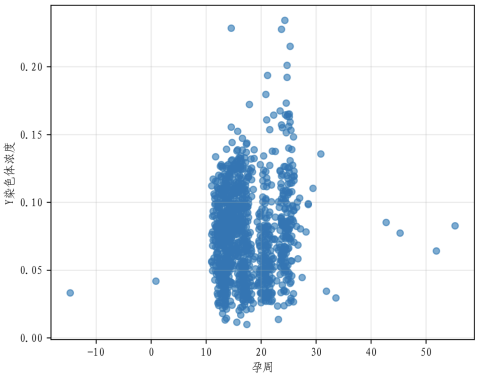

图 1　孕周分布

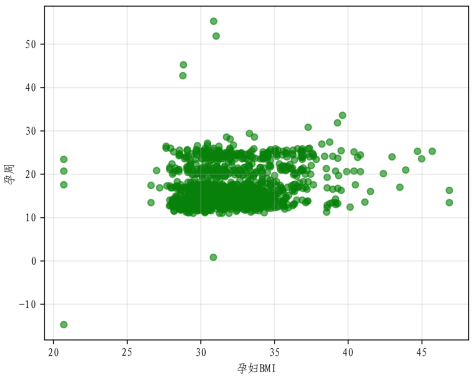

图 2　BMI 分布

由图??分析可得 Y 浓度总体呈现随孕周上升的趋势，但数据离散度较大，高 BMI 样本主要分布在低 Y 浓度区域；Y 浓度相对于 BMI 则表现出下降趋势，尤其在孕周较小时更为明显，且高 BMI 孕妇的 Y 浓度增长速度较慢。

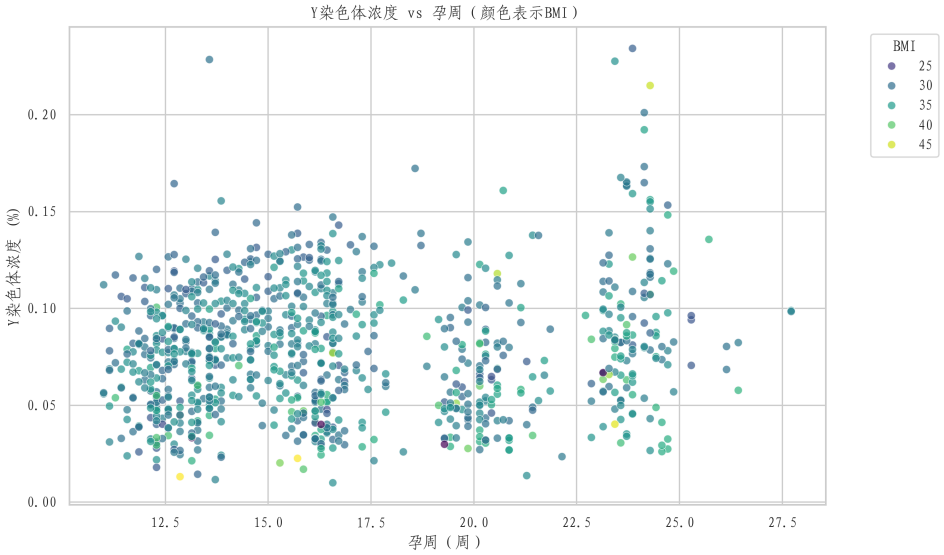

图 3　按 BMI 着色的 Y 浓度 vs 孕周图

计算包含 Y 染色体浓度、孕周、BMI、年龄关键变量的相关系数矩阵后，以热力图形式可视化其相关性，由图 ?? 得出变量的相关性分析

1. Y 染色体浓度与孕周相关系数为 +0.12，呈现弱正相关。这表明随着孕周的增加，Y 染色体浓度有一定程度的上升趋势，但这种关系并不十分强烈。
2. Y 染色体浓度与 BMI 相关系数为 -0.13，呈现弱负相关。说明孕妇的 BMI 越高，Y 染色体浓度可能越低。
3. Y 染色体浓度与孕妇年龄相关系数为 -0.12，可能反映了孕妇的年龄越大，Y 染色体的浓度越低。

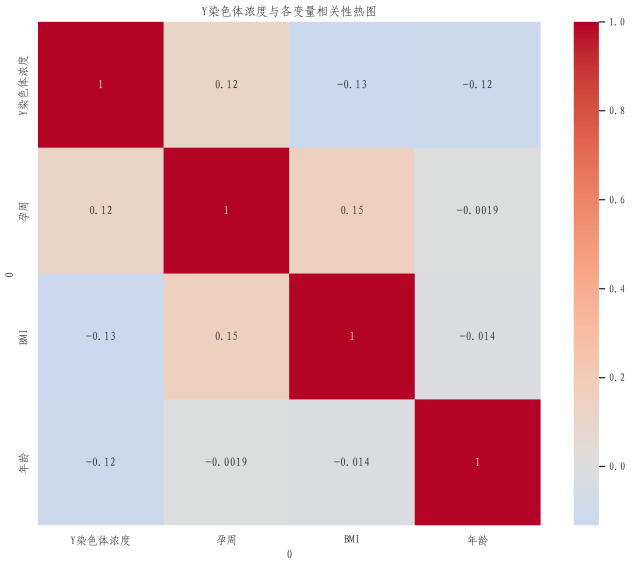

图 4　相关性分析热力图

5.1.3 关系模型

通过相关性分析，我们可以建立以下模型

（一）多元线性回归模型

通过分析相关性矩阵，我们发现 Y 染色体浓度和孕周之间有弱正相关（$r = 0.12$），而 Y 染色体浓度和 BMI 之间有弱负相关（$r = -0.13$）。另外，两者之间没有严重的多重共线性问题（VIF<5），这表明可以用来构建多元线性回归模型。图 **??** 也显示，Y 染色体浓度与孕周、Y 染色体浓度与 BMI 总体上呈现线性变化，没有明显的非线性波动。考虑到题目特别关注孕周和 BMI，我们先用这两个变量建立一个简单的线性模型，这样可以快速抓住基本的关联模式，也为后面更复杂的模型提供一个参考点，适合用来研究 Y 染色体浓度和这些核心指标的关系。该多元线性回归模型的公式为

$$Y = \beta_0 + \beta_1 \cdot GW + \beta_2 \cdot BMI + \epsilon \tag{1}$$

其中 Y 为 Y 染色体浓度，GW 为孕周，BMI 为孕妇的身体质量指数，β_0 为截距，β_1 和 β_2 为回归系数，ϵ 为误差项。对于这个模型，我们通过以下步骤来进行优化求解。

Step 1：基于相关性分析可知，Y 与 GW 相关系数 $r = 0.12$，Y 与 BMI 相关系数 $r = -0.13$，先分别构建单变量模型。

当仅含 GW 时，模型表示为

$$Y = \beta_{01} + \beta_{11} \cdot GW + \epsilon_1 \tag{2}$$

此时残差平方和为

$$SSE_1 = \sum (Y_i \hat{Y}_{i1})^2 \tag{3}$$

当仅含 BMI 时，模型表示为

$$Y = \beta_{02} + \beta_{12} \cdot BMI + \epsilon_2 \tag{4}$$

此时残差平方和为

$$SSE_2 = \sum (Y_i \hat{Y}_{i2})^2 \tag{5}$$

Step 2：因孕周与 BMI 无严重多重共线性（VIF<5），满足线性回归 “无多重共线性假设”，将两个变量整合为多元线性模型，模型表示为

$$Y = \beta_0 + \beta_1 \cdot GW + \beta_2 \cdot BMI + \epsilon \tag{6}$$

此时残差平方和为

$$SSE = \sum (Y_i - (\beta_0 + \beta_1 \cdot GW_i + \beta_2 \cdot BMI_i))^2 \tag{7}$$

Step 3：进行模型参数优化，通过最小化残差平方和来算参数，用最小二乘法得出

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{8}$$

其中 $X = \begin{bmatrix} 1 & GW_1 & BMI_1 \\ 1 & GW_2 & BMI_2 \\ \vdots & \vdots & \vdots \\ 1 & GW_n & BMI_n \end{bmatrix}$ 为设计矩阵，$Y = [Y_1, Y_2, ..., Y_n]^T$ 为因变量向量。

Step 4：基于前一步优化的参数，验证残差正态性（$\epsilon \sim N(0, \sigma^2)$）和方差齐性（$Var(\epsilon_i) = \sigma^2$）均满足假设。F 检验显示模型整体显著（$p = 3.46 \times 10^{-8}$），表明该多元线性回归模型在统计上有效。最终确定的模型参数如表 **??** 所示。从表中提取与 Y 染色体浓度关系的关键信息：

1. 截距为 0.1150（p < 0.001），提供模型的基准值。
2. 孕周的系数为 0.0012（p = 2.20×10^{-5}），显示正向影响显著，每单位孕周平均增加 Y 染色体浓度 0.0012
3. BMI 的系数为 -0.0017（p = 2.90×10^{-6}），显示负向影响显著，每单位 BMI 平均减少 Y 染色体浓度 0.0017