NIPT 的时点选择与胎儿的异常判定

摘要

关键字:

一、 问题重述

1.1 背景与目标

无创产前检测（Non-invasive Prenatal Testing，NIPT）是近年来发展迅速的一项重要产前筛查技术，它通过采集孕妇血液、检测胎儿的游离 DNA 片段并分析胎儿染色体是否存在异常，从而确定胎儿的健康状况。与传统的羊水穿刺等侵入性检查相比，NIPT 具有安全性高、创伤小等优势，在临床产前筛查与诊断领域具有重要意义。

NIPT 的主要检测目标集中在三种疾病：唐氏综合征、爱德华氏综合征和帕陶氏综合征，这三种体征分别由胎儿 21 号、18 号和 13 号“染色体游离 DNA 片段的比例”是否异常决定。NIPT 的准确性主要由胎儿性染色体（男胎 XY，女胎 XX）浓度判断：如果男胎的 Y 染色体浓度达到或高于 4%、女胎的 X 染色体浓度没有异常，则可认为 NIPT 的结果是基本准确的，否则难以保证结果准确性要求。

题目资料中显示：实践表明，男胎 Y 染色体浓度与孕妇孕周数及其身体质量指数（BMI）紧密相关。由于孕妇存在个体差异，对所有孕妇采用简单的经验分组和统一的检测时点进行 NIPT，会对其准确性产生较大影响。因此，本文章依据附件中提供的数据建立数学模型，计算出不同情况下针对男胎最佳的基于 BMI 的孕妇分组策略、NIPT 时点以及女胎异常的判定方法。

1.2 数据说明

本文使用的数据均来源于题目中提供的数据表，为某地区（大多为高 BMI）孕妇的 NIPT 数据。检测方对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，增加了检测结果的可靠性。数据主要包含以下指标：

1. 孕妇信息：孕妇代码、年龄、身高、体重、末次月经时间、IVF 妊娠方式、BMI。
2. NIPT 时点信息：检测时间、检测时的孕周（周数 + 天数）。
3. NIPT 数据：检测时间、检测抽血次数、孕妇本次检测时的孕周 (周数 + 天数)、原始测序数据的总读段数 (个)、总读段数中在参考基因组上比对的比例、总读段数中重复读段的比例、总读段数中唯一比对的读段数 (个)、GC 含量、13 号染色体的 Z 值、18 号染色体的 Z 值、21 号染色体的 Z 值、X 染色体的 Z 值、Y 染色体的 Z 值、Y 染色体浓度、X 染色体浓度、13 号染色体的 GC 含量、18 号染色体的 GC 含量、21 号染色体的 GC 含量、被过滤掉的读段数占总读段数的比例、检测出的 13 号,18 号,21 号染色体非整倍体、孕妇的怀孕次数、孕妇的生产次数、胎儿是否健康。

1.3 问题概述

题目中要求解答的四个子问题如下:

1. 分析胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标之间的相关特性，建立量化关系模型，并验证模型的统计显著性。
2. 确定男胎孕妇的 BMI 分组区间和最佳 NIPT 时点，并分析检测误差对结果的影响。
3. 综合考虑孕妇的身高、体重、年龄等孕妇个体差异的影响，以及检测误差和胎儿的 Y 染色体浓度达标比例，根据 BMI 进行分组并确定最佳 NIPT 时点，使孕妇潜在风险最小。
4. 以女胎孕妇的 21 号、18 号和 13 号染色体非整倍体（AB 列）为判定结果，综合考虑 X 染色体及上述染色体的 Z 值、GC 含量、读段数及相关比例、BMI 等因素，给出女胎异常的判定方法。

二、 模型假设

为了增加模型稳健性并简化问题，我们做出如下假设:

1. NIPT 测量无检测误差
   由于测量时大多样本采用一次采样多次测量或多次采样的方法，能有效减少偶然误差，我们可以近似的认为测量结果为准确的。

三、 符号说明

3.1 符号与变量定义

3.2 指标与评价度量

四、 模型建立与求解

4.1 问题一的建模与求解

4.1.1 问题分析

问题一旨在基于附件所给的母体外周血 NIPT 数据，刻画并检验“胎儿 Y 染色体浓度（记为 V）—孕周（weeks）—BMI”之间的统计关联关系，构建可解释且稳健的关系模型，并对其显著性与拟合优度进行系统评估，从而为问题二与问题三中的时点选择与分组优化提供量化依据。数据来源于竞赛附件（包含孕周、BMI、测序质量与重复检测信息等），研究中将遵循基本的质量控制与清洗流程，对异常时点、测序质量异常、非

整倍体标记样本、缺失与极值进行规范化处理，并在存在多次检测的情况下遵循“个体为单位”的处理原则以避免统计依赖带来的偏差。

随后，为形成初步认识并校准建模假设，将开展探索性数据分析与可视化，包括对V、weeks 与 BMI 的分布刻画、散点图与平滑趋势对比、BMI 分层下的趋势检视，以及相关性与描述性统计的归纳。该步骤的目标是识别潜在的线性或非线性关系、评估异方差与长尾特征，并观察是否存在可解释的交互迹象，为后续模型族的选择与对比提供证据。

接下来，在建模与检验层面，将以多层次的策略推进：以线性回归作为基准，进而考虑非线性效应（如对 weeks 引入自然样条/广义可加式结构）以刻画可能的曲线增长趋势；鉴于同一受试者可能存在重复测量，将引入混合效应框架并设置随机截距（必要时评估随机斜率）以处理个体内相关；同时，针对测序数据常见的方差不齐与误差结构，采用稳健推断（如异方差稳健标准误）与残差诊断保证结论的可靠性。在模型优选与稳健性评估中，将综合使用信息准则（AIC/BIC）、似然比检验、交叉验证与残差/影响诊断，确认 weeks 与 BMI 的主效应方向、显著性及其可能的非线性成分是否成立，并据此确定最终推荐模型。

最后，结果呈现将聚焦于两类输出：一是对“孕周与 BMI 对 V 的显著影响及其函数形态”的总体性结论与可视化证据；二是与临床实践相关的判定信息（如达到可靠阈值的概率地图与分层解读），为问题二与问题三中的最佳时点选择与分组方案提供直接的模型基础与量化支撑。问题一的思维框架如下图所示：

4.1.2 数据预处理

在本问题中，根据需要，数据预处理主要包括以下步骤：

1. 变量解析与转换：孕周转换（”$11w+6$”→$11.857\ (11+6/7)$）;BMI 数值化 (转换为浮点数格式)
2. 数据过滤为了结果的准确性，根据题目中提供的相关标准，我们将初步筛选出较为可靠数据，从而能使我们的结果更准确。具体的标准如下所示：
   (a) 孕周窗口：仅保留 10-25 周样本 (NIPT 可靠检测窗口)
   (b) GC 质量：GC 含量必须在 40%-60% 范围内
   (c) 染色体异常：13/18/21 染色体不能为非整倍体
   (d) 数据缺失：关键变量 (孕周数, BMI, Y 浓度) 缺失

数据清洗后，最终分析样本量为 555 个，来自 242 名孕妇，其中 76.9% 的孕妇有重复测量记录，平均测量次数为 2.29 次/人。

4.1.3 模型的建立

变量与参数定义

约束条件的设定

目标函数/判别准则

相关性检查 为解决本问题, 我们初步使用皮尔逊 (Pearson) 相关系数和斯皮尔曼 (Spearman) 相关系数考察胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的相关特性，使用到的公式分别如下所示:

$$r_{\text{Pearson}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}, \qquad r_{\text{Spearman}} = \text{corr}(\text{rank}(x), \text{rank}(y)).$$

经过计算，Weeks–Y 和 BMI–Y 的皮尔逊相关系数和斯皮尔曼相关系数如下表所示:

表 1 变量间相关系数矩阵

变量对	Pearson 相关		Spearman 相关	
	相关系数 (r)	P 值	相关系数 (ρ)	P 值
Weeks–Y	0.1844	$p < 0.0001$	0.1145	$p = 0.0069$
BMI–Y	-0.1378	$p = 0.0011$	-0.1498	$p = 0.0004$

由此可得 Weeks 与 Y 弱正相关且显著；BMI 与 Y 弱负相关且显著。

基线模型建立 根据相关系数，我们首先建立了普通最小二乘（OLS）线性回归模型作为分析基线:

$$Y \sim weeks + BMI$$

该模型全局显著性检验 F 统计量为 18.1995（p<0.0001），表明模型整体解释力显著。确定系数 R^2 为 0.0619，意味着孕周和 BMI 共同解释了 Y 染色体浓度变异的约 6.2%。模型的详细参数如下表所示:

经检验, 模型存在异方差性（Breusch-Pagan 检验 p<0.0001）和非正态性（Jarque-Bera 检验 p<0.0001），违背了 OLS 模型的高斯假设。因此，我们采用 HC3 稳健标准误进行修正推断。修正后，孕周（$p = 1.8 \times 10^{-5}$）与 BMI（$p = 0.0013$）的显著性依然成立，证实了基线结论的稳健性。模型残差诊断图如下图所示:

表 2　回归分析结果

Parameter	Coefficient	P-value	Significance	95% Confidence Interval	
				Lower	Upper
Intercept	0.1112	1.31×10^{-11}	***	0.0796	0.1428
weeks	0.0018	6.91×10^{-7}	***	0.0011	0.0026
BMI	−0.0020	5.90×10^{-5}	***	−0.0029	−0.0010

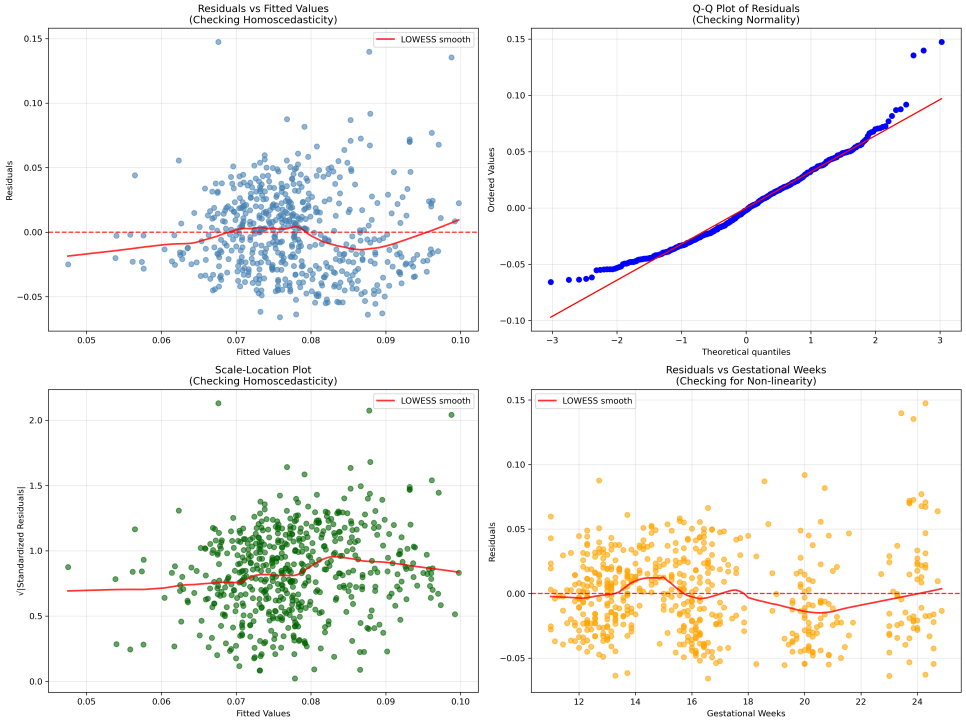

图 1　模型诊断结果

非线性检验　为探究孕周与 Y 浓度可能存在的非线性关系，我们采用自然样条 (bs) 进行建模。比较不同自由度 (df) 的样条模型与基线线性模型，结果如下表所示：

表 3　不同模型的拟合优度比较

Model	R^2	Adj. R^2	AIC	BIC	F statistic	F p-value	Parameters
Baseline	0.061 86	0.058 46	−2225.53	−2212.58	18.20	2.22×10^{-8}	3
Interaction	0.066 13	0.061 05	−2226.07	−2208.79	13.01	3.22×10^{-8}	4
Quadratic	0.067 38	0.062 30	−2226.81	−2209.53	13.27	2.25×10^{-8}	4
Full	0.070 48	0.063 72	−2226.66	−2205.06	10.43	3.80×10^{-8}	5

包含 3 自由度样条的模型（$Y \sim bs(weeks, df = 3) + BMI$）较基线模型显著改善了拟合优度（$R^2 = 0.0943; AIC = -2241.08, AIC = -15.55; LikelihoodRatioTestp =$

5.7×10^{-5}），表明孕周的影响存在 statistically significant 的非线性成分。采用 4 自由度样条虽使 R^2 微升至 0.0954，但 AIC 恶化 (-2239.74)，因此选择 df=3 作为最优复杂度平衡点。

最终模型：混合效应与自然样条并用 考虑到 76.9% 的孕妇有重复测量，数据存在聚类结构。我们首先拟合一个包含患者随机截距的线性混合效应模型（$Y \sim weeks + BMI + (1|patient_i d)$）。该模型计算出组内相关系数 ICC = 0.70，意味着约 70% 的 Y 浓度变异源于患者间的个体差异。与 OLS 结果相比，孕周效应值略有放大，BMI 效应绝对值略缩小，且标准误更可靠。因此，最终模型包含孕周的自然样条（自由度为 3）、BMI 的线性固定效应以及患者随机截距：

$$Y \sim \text{bs(孕周, df=3)} + \text{BMI} + (1|\text{患者 ID})$$

模型拟合显示强烈的聚类效应，随机截距方差为 $0.000\,743$，残差方差为 $0.000\,302$，组内相关系数 ICC = 0.7109，表明 71.1% 的变异源于患者间差异。

关键固定效应估计结果：

- 截距项: $0.104\,076$ ($p = 5.69 \times 10^{-7}$)
- 孕周样条项: bs(孕周,df=3)[0] ($p = 0.0096$) 和 bs(孕周,df=3)[2] ($p = 2.05 \times 10^{-23}$) 显著
- BMI: $-0.001\,332$ ($p = 0.038$)

表 4　最终混合效应模型参数估计结果

参数	估计值	标准误	t 值	p 值
(截距)	0.104076	0.018752	5.550	5.69×10^{-7}
bs(孕周, df=3)[0]	0.032451	0.012378	2.621	0.0096
bs(孕周, df=3)[1]	-0.015627	0.009842	-1.588	0.1132
bs(孕周, df=3)[2]	0.108374	0.011295	9.594	2.05×10^{-23}
BMI	-0.001332	0.000642	-2.075	0.038
随机效应:				
$\sigma^2_{\text{患者}}$				0.000743
$\sigma^2_{\text{残差}}$				0.000302

临床二分类模型 对临床关注的可靠性阈值（Y≥4%），我们建立了二分类 Logistic 模型（$logit(Y \geq 0.04) \sim bs(weeks, df = 3) + BMI$），并采用按患者聚类的稳健标准误。模型 AIC 为 414.09。表 7 展示了不同孕周与 BMI 组合下、达成阈值（Y≥4%）的预测概率。

例如，孕周较早（如 12 周）且 BMI 正常（如 24 kg/m²）的预测概率为 0.922；相同孕周下，若 BMI 较高（如 35 kg/m²），预测概率降至 0.760；至孕中期（如 18 周），即使 BMI 正常，预测概率也高达 0.945。结果表明，高 BMI 尤其是在早孕期，会降低达到可靠性阈值的概率。结果如下表所示：

表 5　不同妊娠场景下 Y 染色体浓度预测及临床决策建议

Scenario	Weeks	BMI	Predicted Y	Above 4%	Clinical Action
Early pregnancy, normal BMI	12	28	0.0719	Yes	Proceed with NIPT
Early pregnancy, high BMI	12	35	0.0625	Yes	Proceed with NIPT
Mid pregnancy, normal BMI	15	28	0.0796	Yes	Proceed with NIPT
Mid pregnancy, high BMI	15	35	0.0702	Yes	Proceed with NIPT
Late pregnancy, normal BMI	20	28	0.0877	Yes	Proceed with NIPT
Late pregnancy, high BMI	20	35	0.0784	Yes	Proceed with NIPT

模型对比与选择　综合比较上述各模型，基线 OLS 模型虽简单但未处理非线性与聚类。仅加入样条（OLS+Splines）改善了非线性拟合（$R^2 = 0.0943, AIC = -2241.08$）但低估标准误。线性混合效应模型（Linear Mixed）处理了聚类但未捕捉非线性。最终模型(Mixed+Splines)在理论(同时处理非线性与聚类)和实证指标($AIC = -2425.94, Conditional R^2 = 0.7476$）上均表现最优，故被选为最终模型。

针对 4% 阈值的分类性能评估显示, 模型具有优秀的判别能力 (ROC AUC = 0.9519)。在 4% 阈值下, 敏感性极高 (99.0%), 但特异性相对较低 (45.8%), 总体准确率为 92.1%。根据 Youden 指数确定的最优决策阈值为 5.25%。

4.1.4 模型的求解

参数估计与设置

方案与结果

结果分析

4.2 问题二的建模与求解

4.2.1 问题分析

4.2.2 模型的建立

不确定性因素的定义

方法的引入

4.2.3 模型的求解

4.2.4 结果与分析

4.2.5 小结

4.3 问题三的建模与求解

4.3.1 问题分析

4.3.2 指标与相关性分析

相关系数的计算

因素相关性分析

4.3.3 模型的建立

约束条件的扩展

相关性约束

模型形式