

NIPT 的时点选择与胎儿的异常判定

摘要

关键字： NIPT BMI 分组 生存分析 Logistic 回归 风险最小化

一、问题重述

1.1 背景与目标

1.2 数据说明

1.3 问题概述（四个子问题）

二、模型假设

2.1 基本假设

2.2 误差与稳健性相关假设

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 问题分析

4.2 问题一的建模与求解

4.2.1 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：

4.2.2 数据预处理

4.2.3 指标与可视化分析

指标分布可视化

相关性分析可视化

关键指标对比可视化

4.2.4 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

4.2.5 模型的求解

参数估计与设置

方案与结果

结果分析

4.3 问题二的建模与求解

4.3.1 问题分析

4.3.2 模型的建立

不确定性因素的定义

方法的引入

4.3.3 模型的求解

4.3.4 结果与分析

4.3.5 小结

4.4 问题三的建模与求解

4.4.1 问题分析

4.4.2 指标与相关性分析

相关系数的计算

因素相关性分析

4.4.3 模型的建立

约束条件的扩展

相关性约束

模型形式

4.4.4 结果与分析

4.4.5 小结

4.5 问题四的建模与求解

4.5.1 问题分析

4.5.2 特征构造与数据处理

4.5.3 模型的建立

4.5.4 模型训练与验证

4.5.5 模型评价与对比

4.5.6 小结

五、 灵敏度分析

5.1 基于鲁棒优化的灵敏度分析

5.2 基于动态调参的灵敏度分析

六、 模型评价与推广

6.1 模型的优点

6.2 模型的缺点

6.3 模型的推广

七、 结论与展望

参考文献