

数学建模比赛

摘要

本文针对 NIPT（无创产前检测）技术中最佳检测时点选择与胎儿异常判定的问题，通过建立数学模型，对孕妇进行合理分组并确定最佳检测策略，以最小化潜在风险并提高检测准确性。我们综合利用了数理统计等方法，系统地解决了题目提出的四个问题。

针对问题一，首先处理表格中的数据，把检测次数不足四次的孕妇代码剔除并把检测孕周换算成天数，方便后使之后的线性回归计算更加可靠。之后剩下的每个人的 y 染色体浓度与检测孕周（天数）进行线性回归计算方程与 R^2 ，考察到整体 y 染色体浓度与检测孕周的线性拟合程度都很高。之后对于这些人，求出检测的时间中 BMI 增长速率。将每个人的线性回归程斜率与进行相关性分析，得到了一套比较显著可靠的关系模型。最后引入斯皮尔曼系数进行检验

针对问题二

针对问题三

针对问题四

关键词：线性回归，斯皮尔曼系数

一 问题重述

1.1 问题背景

NIPT (Non-invasive Prenatal Testing, 无创产前检测) 作为一项革命性的产前筛查技术, 通过采集孕妇外周血中的胎儿游离 DNA 进行测序分析, 能够有效评估胎儿常见染色体非整倍体异常的风险 [1]。该技术具有无创、安全、准确性高等特点, 已成为临床产前筛查的重要手段。

在实际临床应用中, NIPT 检测的准确性受到多种因素影响, 其中胎儿游离 DNA 浓度 (特别是 Y 染色体浓度对于男胎) 是关键因素之一。临床实践表明, 胎儿 Y 染色体浓度与孕妇孕周数及身体质量指数 (BMI) 存在显著相关性 [2]。同时, 检测时机的选择对于尽早发现胎儿异常、降低临床风险至关重要: 早期发现 (12 周以内) 风险较低, 中期发现 (13-27 周) 风险较高, 而晚期发现 (28 周以后) 风险极高。

目前临床通常根据孕妇 BMI 值进行简单分组并确定统一的检测时点, 但这种方法未能充分考虑孕妇年龄、体重等个体差异, 可能导致部分孕妇错过最佳检测时机, 增加临床风险。因此, 需要建立更加科学、个性化的 NIPT 时点选择模型, 为不同特征的孕妇群体制定最优检测策略。

1.2 问题要求

附件提供了某地区 (大多为高 BMI) 孕妇的 NIPT 检测数据, 包括孕妇年龄、BMI、孕周数、胎儿染色体浓度、Z 值、GC 含量、读段数等相关指标。现需要根据这些数据建立数学模型, 解决以下问题:

问题一: 基于附件中的数据, 分析胎儿 Y 染色体浓度与孕妇孕周数、BMI 等指标的相关特性, 建立合适的数学模型描述它们之间的关系, 并对模型的显著性进行统计检验。

问题二: 临床证明男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度达标时间 (浓度 4% 的最早时间) 的主要因素。请对男胎孕妇的 BMI 进行合理分组, 确定每组的 BMI 区间和最佳 NIPT 检测时点, 使得孕妇的潜在风险最小, 并分析检测误差对结果的影响。

问题三: 综合考虑体重、年龄等多种因素对男胎 Y 染色体浓度达标时间的影响, 同时考虑检测误差和胎儿 Y 染色体浓度达标比例, 根据男胎孕妇的 BMI 进行合理分组, 确定每组的最佳 NIPT 检测时点, 使孕妇潜在风险最小, 并分析检测误差对结果的影响。

问题四: 针对女胎异常的判定问题, 以女胎孕妇的 21 号、18 号和 13 号染色体非整倍体为判定结果, 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, 建立女胎异常的判定模型和方法。

二 问题分析

2.1 问题一的分析

针对问题一, 我们需要先对附件中的数据进行预处理, 为之后对每个人进行线性回归分析, 剔除掉同一个人检测次数小于等于三的孕妇代码, 使拟合结果更加可靠。之后对于每一个人, 以 y 染色体为 y

轴，检测孕周（天数）为 x 轴进行线性回归计算，并求 R^2 确定拟合程度。之后对与每个孕妇代码求出 BMI 增长率，对比斜率与 BMI 增长率，检测其相关性。

问题一：

问题二：

问题三：

三 模型假设

1. 假设附件所提供的孕妇 NIPT 检测数据真实、准确，且数据样本足以反映胎儿染色体浓度与孕妇孕周、BMI 等指标间的统计规律。
2. 假设题目中给出的 Y 染色体浓度 4% 的临界值是可靠且普适的。
3. 假设具有相似特征（如处于同一 BMI 区间）的孕妇群体，其胎儿 Y 染色体浓度的增长规律和达标时间具有相似的统计特征。
4. 假设题目所提供的特征（包括但不限于 X、21、18、13 号染色体的 Z 值、GC 含量、读段数比例及孕妇 BMI 等）包含了足以有效判别女胎染色体是否异常的信息。

四 符号说明

符号	说明
k_i	孕妇代码为 i 的个体 Y 染色体浓度与孕周线性回归方程斜率
c_{ij}	孕妇代码为 i 的个体第 j 次检测得到的 Y 染色体浓度
t_{ij}	孕妇代码为 i 的个体第 j 次检测的检测孕周（转化为天数）
b_i	孕妇代码为 i 的个体 BMI 增长率
\bar{b}	所有孕妇个体 BMI 增长率的平均数

五 问题一模型建立求解

5.1 模型建立思路

问题一要求分析胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标的相关特性。为精确刻画个体增长规律并避免群体平均带来的偏差，我们决定采用基于个体时间序列的线性拟合方法。核心思路是：首先对数据进行清洗，保证每个个体的数据点足以支持回归分析；然后为每一位符合条件的孕妇单独建立 Y

染色体浓度随孕周（天数）变化的线性回归模型，以拟合斜率量化其增长速率；最后，将所有个体的增长速率（斜率）与其对应的 BMI 指标进行相关性分析，从而揭示 BMI 对 Y 染色体浓度增长的影响。

5.2 数据预处理

附件中的数据存在个别孕妇检测次数过少的情况，这会导致线性回归结果不可靠。为确保模型稳定性，我们设定了数据筛选条件：仅保留检测次数大于 3 次的孕妇数据。

对于计算出来的结果，我们剔除掉以下拟合直线：

1. R^2 小于 0.5。此时拟合效果差。

2. 斜率小于 0 此时 Y 染色体。此时浓度呈现随时间下降趋势，而经过收集资料，我们发现孕妇体内胎儿 Y 染色体浓度的正常趋势是：孕早期（10-12 周）从无法检出到逐步升高 → 孕中期（12-22 周）稳步升至峰值 → 孕晚期（> 22 周）进入稳定平台期（允许轻微波动，但始终可检出）。

经过预处理，原始数据中共包含 261 名孕妇的记录，其中符合要求的有效孕妇样本为 178 名。

5.3 个体线性回归模型

对于每一位有效孕妇个体 i ，以其多次检测的孕周（转换为天数 t_{ij} ）为自变量，对应的 Y 染色体浓度（ c_{ij} ）为因变量，建立一元线性回归模型：

$$c_{ij} = k_i * t_{ij} + b_i + \epsilon_{ij} \quad (1)$$

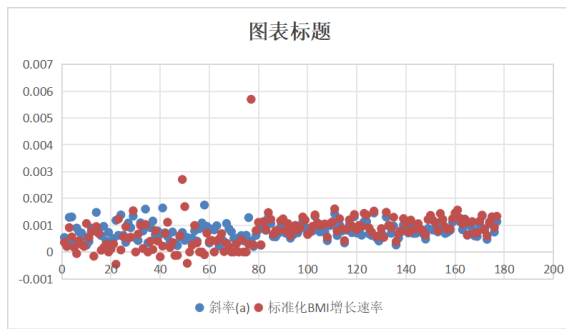
其中： k_i 表示第 i 位孕妇的胎儿 Y 染色体浓度的日增长率，这是我们关注的核心参数； b_i 为截距项； ϵ_{ij} 为随机误差项。

我们采用最小二乘法进行参数估计，并计算确定系数 R_i^2 以评估拟合优度

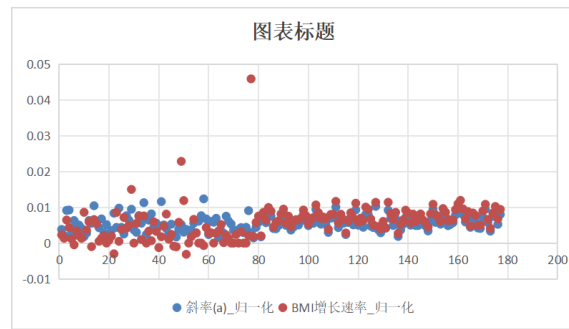
5.4 BMI 与增长速率的全局相关性分析

为探究 BMI 是否导致了上述增长速率的差异，我们计算了每位孕妇的 BMI 增长率个体 Y 染色体浓度增长率 k_i 的相关性。

我们绘制了标准化 BMI 增长速率与斜率 k_i 的散点图和对两者都进行归一化之后的散点图，如下



(a) 标准化 BMI 增长速率与斜率的散点图



(b) 归一化 BMI 增长速率与归一化斜率的散点图

我们可以看到这两组点的吻合程度极高，由此得出结论：BMI 增长率个体 Y 染色体浓度增长率有较高关联性。

之后，我们采用 **Pearson 相关系数** r 来衡量二者之间的线性相关强度，并计算其显著性 p 值以判断该相关性是否由随机因素引起。计算公式如下：**Pearson 相关系数计算公式**：

$$r = \frac{\sum_{i=1}^n (b_i - \bar{b})(k_i - \bar{k})}{\sqrt{\sum_{i=1}^n (b_i - \bar{b})^2} \sqrt{\sum_{i=1}^n (k_i - \bar{k})^2}} \quad (2)$$

其中， n 为有效孕妇样本数， b_i 和 k_i 分别为第 i 位孕妇的初始 BMI 和 Y 染色体浓度日增长率， \bar{b} 和 \bar{k} 分别为它们的样本均值。

显著性检验（t 检验）统计量及 p 值计算公式：

为检验相关系数的显著性，构建 t 统计量：

$$t = r \sqrt{\frac{n-2}{1-r^2}} \quad (3)$$

我们通过上述公式计算得到 r 值为 0.2795， p 值为 0.000165，满足 $p < 0.001$ 。因此，基于统计结果，我们有极大把握认为两者具有显著相关性。这表明，**BMI 更高的孕妇，其胎儿 Y 染色体浓度随孕周增长的速度更快**。这一结论与临床经验相符，为临床上进行个性化 NIPT 时点预测提供了重要的数学依据。