赛区评阅编号（由赛区组委会填写）：

2025 高教社杯全国大学生数学建模竞赛

承　诺　书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（以下简称“竞赛章程和参赛规则”，可从 http://www.mcm.edu.cn 下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：C

我们的报名参赛队号（12 位数字全国统一编号）：4321

参赛学校（完整的学校全称，不含院系名）：中山大学

参赛队员 (打印并签名) ：1. 陈昊蔚

2. 李可乐

3. 蔡佳陆

指导教师或指导教师组负责人 (打印并签名)：指导老师

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期: 2025 年 9 月 5 日

（请勿改动此页内容和格式。此承诺书打印签名后作为纸质论文的封面，注意电子版论文中不得出现此页。以上内容请仔细核对，如填写错误，论文可能被取消评奖资格。）

赛区评阅编号：
（由赛区填写）_____

全国评阅编号：
（全国组委会填写）_____

2025 高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

评阅人						
备注						

送全国评阅统一编号:
（赛区组委会填写）

（请勿改动此页内容和格式。此编号专用页仅供赛区和全国评阅使用，参赛队打印后装订到纸质论文的第二页上。注意电子版论文中不得出现此页。）

论文题目

摘要

摘要的具体内容。

关键字： Pearson 检验　Spearman 检验　多元非线性回归分析　粒子群算法　NIPT 技术　胎儿染色体异常

一、 问题重述

本节旨在提取题目的关键信息，全面概括关于 NIPT 时点选择与胎儿异常判定的背景，并进一步明确根据孕妇的 BMI、孕周数、孕情等个体差异，推断出既能确保准确性、又能尽量降低治疗窗口期缩短的风险的最佳 NIPT 时点以及针对女胎异常的判定方法的现实要求，从而更加清晰地把握问题的核心要点。

1.1 问题背景

进入新时代，为了响应国家“晚婚晚育、少生优生”的号召，切实提高人口素质，许多家庭选择在较晚的年龄生育子女。然而，随着高龄产妇比例的增加，胎儿染色体异常的风险也随之上升。因此，如何在孕期早期准确筛查出胎儿染色体异常，成为了产前诊断领域亟需解决的重要问题。

NIPT（Non-Invasive Prenatal Test，无创产前检测）是一种产前检测技术，仅需对孕妇采集血样就可检测出其中的胎儿游离 DNA 片段，并分析胎儿染色体是否存在异常（例如唐氏综合征患者的 21 号染色体数量异常），从而在早期就可掌握胎儿的健康状况。NIPT 技术可以有效筛查唐氏综合征、爱德华氏综合征和帕陶氏综合征这三大染色体异常疾病，准确率远超先前其他方法。此外，NIPT 技术无需侵入性操作，避免了传统产前诊断方法可能带来的流产风险和对胎儿可能造成的伤害，因而被广泛应用于临床实践中。

我们本次研究的核心任务，便是基于一批孕妇的 NIPT 检测数据，构建有效的数学模型。我们希望能够借助数学模型分析胎儿 Y 染色体浓度和孕妇孕情的关系、不同 BMI 孕妇的最佳 NIPT 时点以及针对无 Y 染色体的女胎的异常判定方法等一系列问题。因此，如何利用现代数据分析与数学建模技术，排除或修正这些数据中潜在的干扰，准确地对上述问题完成模型建立与求解，成为了一个兼具医学意义与数据科学挑战的交叉学科课题。

1.2 基本问题

附件是某地区 BMI 偏高孕妇的 NIPT 检测数据，包含了孕妇的基本信息、孕情信息以及 NIPT 检测结果等内容。为了依据相关数据完成对更多孕妇的 NIPT 时点判断以及针对女胎的异常判定工作，现需要结合这些数据和已知条件，建立数学模型，分析以下问题：

问题一：依据男胎检测数据，分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标之间的相关特性，对有相关性的指标进行筛选，并建立数学模型，描述胎儿 Y 染色体

浓度与孕妇孕周数和 BMI 等指标之间的关系。之后，再检验该模型的显著性。

问题二：依据男胎检测数据，建立数学模型，对男胎孕妇的 BMI 进行分组，分析不同组别男胎孕妇的最佳 NIPT 时点，并分析检测误差对判断最佳 NIPT 时点的影响。

问题三：依据男胎检测数据，综合考虑孕妇的身高、体重、年龄、生育次数等多种因素的影响，并结合检测误差以及胎儿 Y 染色体浓度达标比例，对孕妇的 BMI 进行再次分组，分析各组别孕妇的最佳 NIPT 时点，并分析检测误差对判断最佳 NIPT 时点的影响。

问题四：依据女胎检测数据，建立数学模型，综合考虑 X 染色体和其他检测染色体的 NIPT 结果 Z 值、GC 含量、读数段及相关比例、BMI 等因素，分析女胎孕妇的 21 号、18 号、13 号染色体非整倍体结果，给出判断女胎是否异常的方法。

二、 问题分析

2.1 问题一分析

问题一要求我们依据男胎检测数据，分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标之间的相关特性。首先，我们需要对数据进行预处理，使得数据尽可能完整、合理。然后，我们可以使用统计分析方法，如 Pearson 检验和 Spearman 检验的相关系数计算和多元非线性回归分析，来探索胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标之间的关系。通过建立数学模型，我们可以量化这些关系，并检验模型的显著性，以确保其可靠性。

具体步骤包括：

- 数据预处理：清洗数据，处理缺失值和异常值。
- 相关性分析：计算胎儿 Y 染色体浓度与孕妇孕周数和 BMI 等指标的相关系数。
- 模型建立：选择合适的回归模型（如多项式回归等）来描述这些关系。
- 模型检验：使用合适的指标（t 检验、RMSE、调整后 R^2 等）来评估模型的显著性和拟合优度。

2.2 问题二分析

问题二指出，影响 Y 染色体浓度检测结果的主要因素是孕妇的 BMI。基于此，我们需要对男胎孕妇的 BMI 进行区间划分，对于每个组别，分析其最佳 NIPT 时点。由于检测时间过晚可能会导致错过最佳治疗窗口期，因此为了满足题意中的“尽量降低治疗窗口期缩短的风险”，我们应当在确保检测的准确性，也就是 Y 染色体浓度大于等于 4% 的前提下，选择尽可能早的 NIPT 时点。我们将各组别的最佳 NIPT 时点乘以该组别的权重，求和作为总的检测风险值，并以此作为优化目标，使用粒子群算法等优化方法来求解各组别的最佳 NIPT 时点。

三、 模型假设

结合题意和上述对问题的分析，为了合理简化模型的建立与求解过程，我们提出了以下假设：

- **假设 1**：数据中的测量误差是随机分布的，不会系统性地偏向某一方向。
- **假设 2**：孕妇的身体状况和生活习惯等非 NIPT 检测指标在研究期间保持相对稳定，不会对胎儿 Y 染色体浓度产生显著影响。
- **假设 3**：不同孕妇之间的个体差异可以通过统计方法进行控制和调整。
- **假设 4**：胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标之间的关系可以通过多元非线性回归模型进行描述。
- **假设 5**：在问题二和问题三中，孕妇的 BMI 分组是合理且具有代表性的。考虑到各组别中 BMI 的分布不一定足够均匀，假定每个组别的 BMI 中位数对应的最佳 NIPT 时点能够代表该组别的最佳 NIPT 时点。
- **假设 6**：孕妇的潜在风险仅取决于 NIPT 时点的选择，而不受其他外部因素的影响。

四、 符号说明

概念/符号	意义
BMI	身体质量指数，BMI = 体重（kg）/身高^2（m^2）
Z 值	数据经过 Z-Score 标准化后得到的数值
Y	胎儿 Y 染色体浓度
X_1	孕周数
X_2	孕妇 BMI
t_i	第 i 组别的最佳 NIPT 时点
w_i	第 i 组别的权重
$f(t_1, t_2, t_3, t_4, t_5)$	总的检测风险值
N	总的孕妇人数
n_i	第 i 组别的孕妇人数

五、 模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 数据预处理

在进行建模之前，我们应当对附件原始数据进行合理的预处理，包括以下几个方面：

缺失值处理：对于缺失的数据，采用均值填补、插值法或删除含有缺失值的样本等方法进行处理。

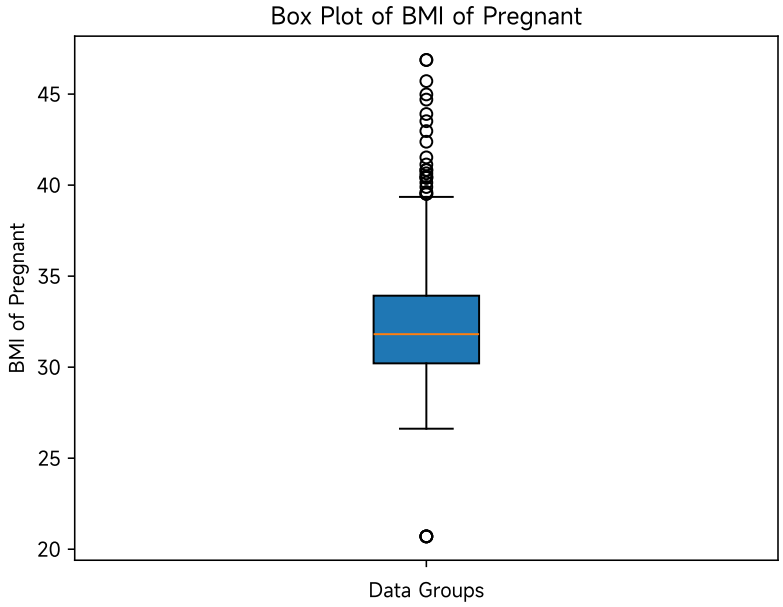

图 **1**　男胎孕妇 **BMI** 箱线图

异常值检测：使用箱线图识别 BMI 等变量中的异常值，并根据具体情况决定是否剔除或修正这些异常值。使用 Python 绘制 BMI 的箱线图，发现在上须线以上有大量异常点，在下须线以下存在一个异常点，如图 1 所示。

经过使用身高、体重的原始数据重新计算 BMI，我们发现异常值与重新计算值的相对误差不超过 1‰。然而，由于题意中描述的数据人群主要是 BMI 较大的孕妇，因此 BMI 为 25 以下的，我们将其视为离群值，予以舍弃。对于上述在上须线以上的异常值，我们则选择保留。

此外，对于 Y 染色体浓度超过 20% 的异常值，显然已经不符合生物学常识，可能是检测错误，因此我们选择剔除这些异常值。

纵向数据整理：对于“一次抽血多次检验”的数据，取其中位数作为该孕妇的最终检测结果，以减少偶然误差的影响。

5.1.2 相关性检验

为了分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标之间的相关特性，我们首先计算这些变量之间的相关系数。

由于 BMI 和孕周数均为连续变量，我们可以使用 Pearson 相关系数来衡量它们与 Y 染色体浓度之间的线性关系。同时考虑到该关系也有可能是非线性的，因此我们也计算了 Spearman 秩相关系数，以便交互验证。计算得到的两种相关系数热力图如图 2 所示。

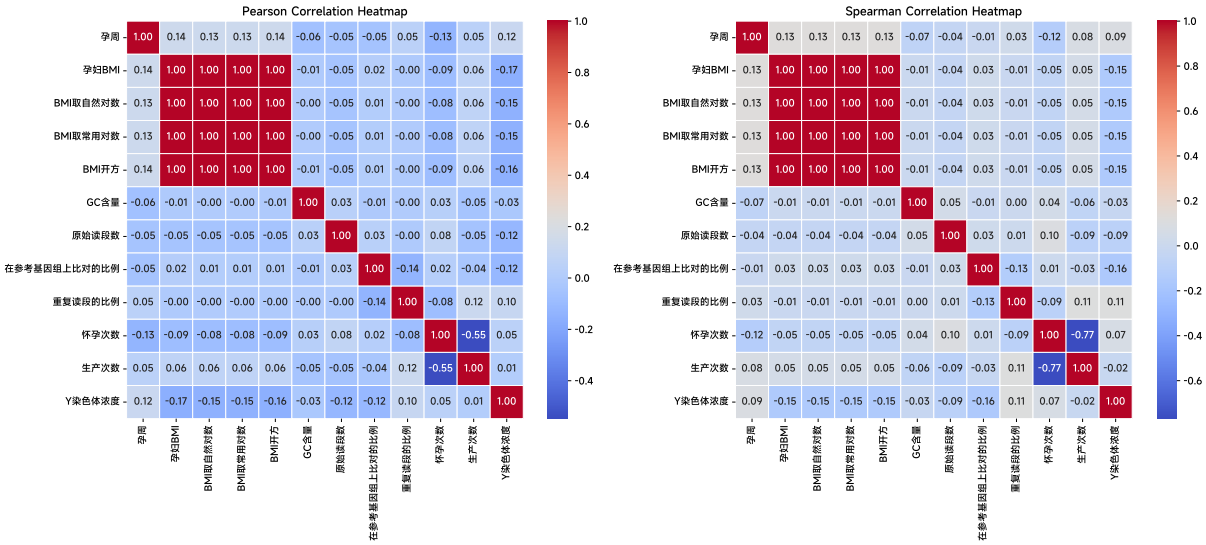

图 2　**Pearson** 与 **Spearman** 相关系数热力图

由图 3 可见，与 Y 染色体浓度的相关性检验满足 P 值法的因素有：孕周、BMI、原始读段数、在参考基因组上比对的比例、重复读段的比例等。然而，显然“原始读段数”“在参考基因组上比对的比例”和“重复读段的比例”仅仅是判断 NIPT 检测过程中样本量是否足够大、是否有重复测量的技术指标，并不具备生物学意义，因此我们将其排除在影响 Y 染色体浓度的因素之外。对于剩下的因素，可以发现 Y 染色体浓度与孕周数和孕妇 BMI 的 P 值均小于 0.001，表明两者与 Y 染色体浓度之间均存在显著的相关性。同时又可以发现 Y 染色体浓度与孕周数的 Pearson 相关系数和 Spearman 相关系数均为接近 0 的正数，表明两者之间存在正向的关系，但是线性相关性很弱；同理，Y 染色体浓度与孕妇 BMI 的 Pearson 相关系数和 Spearman 相关系数均为接近 0 的负数，表明两者之间存在负向关系，但是线性相关性也很弱。

综合上述分析结果，我们决定在后续模型中同时考虑孕周数和 BMI 对 Y 染色体浓度的影响，并重点考虑两者之间可能存在的非线性关系。

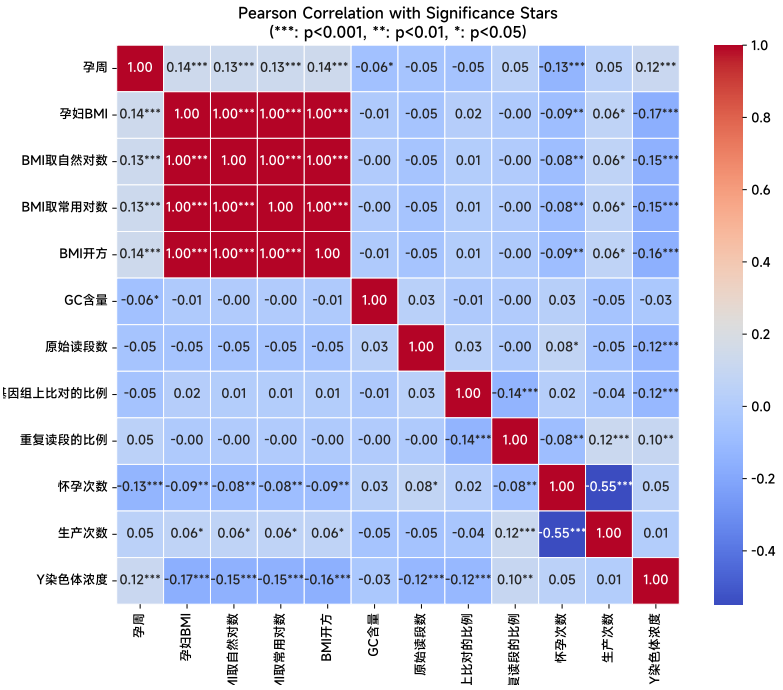

图 **3**　**Pearson** 相关系数和 **P** 值热力图

5.1.3 回归模型建立与显著性的检验

由于 Y 染色体浓度与孕周数和 BMI 之间的相关性较弱，我们决定采用多元线性回归模型以及多元非线性回归模型来先后尝试描述它们之间的关系，后者主要包括包含交互项的二阶多项式回归模型、包含交互项的三阶多项式回归模型和双对数模型。其函数模型如表 1 所示。其中，Y 表示 Y 染色体浓度，X_1 表示孕周数，X_2 表示 BMI，$\beta_i(i=1,2,\dots)$ 和 d 均为待定系数。

模型名称	定义公式
包含交互项的二阶多项式回归模型	$Y=\beta_1X_1^2+\beta_2X_2^2+\beta_3X_1X_2+\beta_4X_1+\beta_5X_2+d$
包含交互项的三阶多项式回归模型	$Y=\beta_1X_1^3+\beta_2X_2^3+\beta_3X_1^2X_2+\beta_4X_1X_2^2+\beta_5X_1^2+\beta_5X_1^2+\beta_6X_2^2+\beta_7X_1X_2+\beta_8X_1+\beta_9X_2+d$
双对数模型	$\ln Y=\beta_1+\beta_2\ln X_1+\beta_3\ln X_2$

表 **1**　问题一的回归模型建立

经过计算，我们得到了各个模型的待定系数，并使用调整后的 R^2、RMSE 和 AIC 等指标对各个模型的拟合优度进行了评估，同时为了检验该模型的显著性，还做了对模型整体的 F 检验，结果如表 2 所示。

由表 2 可见，尽管包含交互项的三阶多项式回归模型在各项指标上均优于其他模型，但由于其 F 检验的 P 值过大，因而我们选择相对更加稳健的双对数模型作为最终的回归模型。经过计算，我们得到了该模型的待定系数为 $\beta_1=0.8528$，$\beta_2=0.2305$，

模型名称	RMSE	AIC	BIC	调整后 R^2	F 检验的 P 值
二阶多项式回归模型	0.0307	-7050.1141	-7020.5841	0.0650	1.0000
三阶多项式回归模型	0.0303	-7070.2017	-7020.9851	0.0906	1.0000
双对数模型	0.0319	-6978.2983	-6963.5333		0

表 **2**　问题一的回归模型指标比较

β_3 = -0.9045。所以问题一的回归模型为

$$\ln Y = 0.9004 + 0.2299 \ln X_1 - 0.9195 \ln X_2 \tag{1}$$

该函数的拟合效果如图 4 所示。

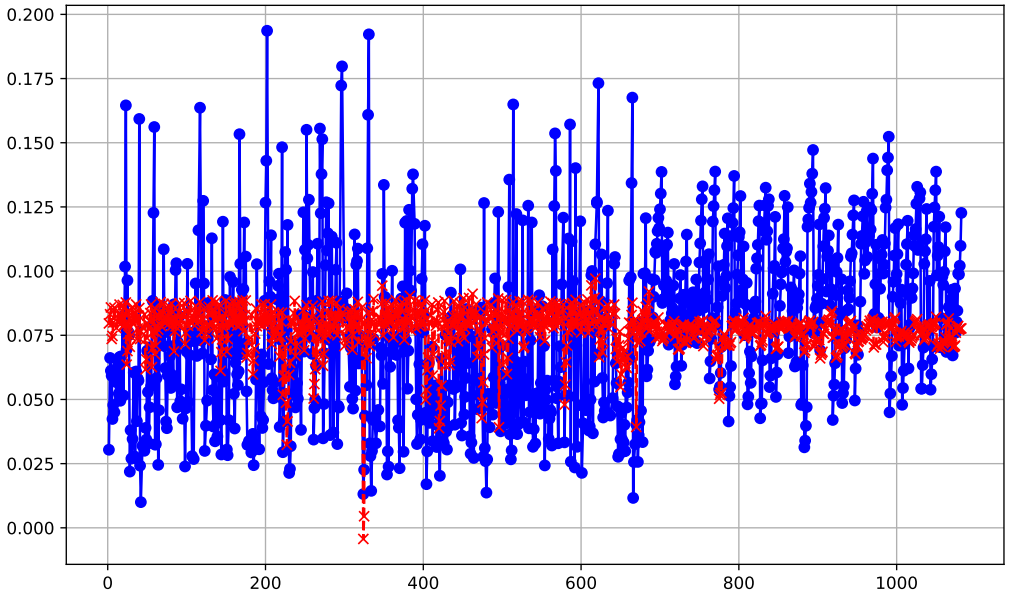

图 **4**　问题一的回归模型拟合效果

5.2 问题二的模型建立与求解

5.2.1 确立优化模型

参考题干中提到的分组举例，我们先假定可以将 BMI 划分为 5 组。由于在达到 Y 染色体浓度大于等于 4%，也即 NIPT 检测准确率有所保障的前提下，孕妇做 NIPT 检测的时间越早则对孕妇和胎儿的后续检测和治疗风险越小，因此我们利用“时间早”和“风险小”的一致性，将风险函数定义为各组别的最佳 NIPT 时点乘以该组别的权重之

和，也即

$$f(t_1, t_2, t_3, t_4, t_5) = \sum_{i=1}^{5} w_i \cdot t_i \tag{2}$$

其中，$w_i = \dfrac{n_i}{N}$。同时，为了确保 NIPT 检测的准确性，我们还需要添加约束条件如下：

- $t_i \geq 9$，即 NIPT 检测时点不早于孕 9 周（对于绝大多数数据样本，9 周以前的 Y 染色体浓度均小于 4%）；
- $t_i \leq 24$，即 NIPT 检测时点不晚于孕 28 周（由题意知，28 周以后发现异常风险极高）；
- $n_i \geq 20$，即每个组别的孕妇人数不少于 20 人。

综上所述，我们可以将问题二的优化模型表述为：

$$\min f(t_1, t_2, t_3, t_4, t_5) = \sum_{i=1}^{5} w_i \cdot t_i \tag{3}$$

$$\text{s.t.} \begin{cases} 9 \leq t_i \leq 24 & , i = 1, 2, \ldots, 5 \\ n_i \geq 20 & , i = 1, 2, \ldots, 5 \end{cases}$$

5.2.2 粒子群算法求解 BMI 分组与各组别最佳 NIPT 时点

得到了上述目标函数和约束条件后，我们便可以使用粒子群算法来求解各组别的最佳 NIPT 时点。

5.3 问题三的模型建立与求解

5.4 问题四的模型建立与求解

六、 总结

参考文献

[1] ....

附录的内容。