

DeepSeek v3.1

1. 资料查找阶段，使用 AI 寻找 NIPT 与 BMI 关系的相关文献

帮我寻找和无创产前检测时间和BMI的关系有关的文献

已深度思考 (用时 20 秒) >

您好！很高兴能协助您进行数字建模竞赛的文献调研。针对“无创产前检测（NIPT）时间选择与孕妇BMI关系”这一主题，该领域确实已有较为明确的研究共识。

核心结论是：**孕妇的BMI越高，其血液中的胎儿游离DNA (fetal fraction, FF) 比例通常越低**。因此，为了达到成功检测所需的FF阈值（例如对男胎Y染色体检测的4%），高BMI孕妇可能需要等待更长的孕周（即推迟检测时间）。

以下是一些关键文献和检索策略，希望能为您的模型提供理论支持。

一、核心推荐文献（可直接用于支撑模型）

您可以通过【Google Scholar】、【PubMed】 (<https://pubmed.ncbi.nlm.nih.gov/>) 或【中国知网 CNKI】等数据库，使用下述标题或DOI进行查找。

- 文献标题:** Analysis of fetal fraction in maternal plasma cell-free DNA and its correlation with maternal characteristics in Chinese women
 - 作者: Y. Liu, et al.
 - 期刊: *Prenatal Diagnosis*
 - 发表年份: 2021年左右
 - 摘要: 这篇研究对中国人群进行了分析，明确展示了胎儿游离DNA比例与孕妇体重、孕周、BMI等指标的相关性。**非常适合作为您问题1中建立关系模型的基础**。文中很可能提供了具体的统计数据和相关性系数。
- 文献标题:** The impact of maternal plasma DNA fetal fraction on next generation sequencing tests for common fetal aneuploidies

Deepseek 推荐的文献为我们寻找与 BMI 有关的因素变量提供了一定的参考，我们在文章中对其相关性进行了检验。

2. 构建模型阶段，询问 AI 关系模型的相关解释，了解相关信息

什么叫相应的关系模型

已深度思考 (用时 18 秒) >

好的，这是一个非常核心的问题。在数学建模和数据分析的语境中，“**相应的关系模型**”指的是：

用一个或一组数学公式，来定量地描述一个或多个自变量（解释变量）与因变量（被预测变量）之间关系的数学模型。

对于本题的第一问，具体来说就是：**建立一个以“孕妇孕周数”和“孕妇BMI”为自变量，以“胎儿Y染色体浓度”为因变量的数学方程**，用以揭示和量化它们之间的相关特性。

“关系模型”的核心组成部分：

- 因变量 (Dependent Variable)：**我们感兴趣并试图预测或解释的变量。
 - 在本问题中：**胎儿Y染色体浓度** (V列)。
- 自变量 (Independent Variables)：**用来解释因变量变化原因的变量，也称为解释变量。
 - 在本问题中：**孕妇孕周数** (J列)、**孕妇BMI** (K列)。在更复杂的模型中，还可以加入年龄、体重等。
- 函数形式 (Functional Form)：**自变量和因变量之间关系的数学表达形式。这是“关系模型”的核心。
- 模型参数 (Parameters)：**函数中的系数，需要通过数据来估计。这些参数的大小和符号揭示了关系的强度和方向。

针对第一问可能的“关系模型”示例：

- 多元线性回归模型 (最简单直接)**

Deepseek 为我们解释了关系模型的含义，我们意识到构建关系模型的核心是为自变量和因变量构建合适的函数关系，以此来解释他们之间的联系。

同时他也给予了一些建模建议，我们通过分析发现并不适用本题目，最终决定不采纳。

3. 构建模型阶段，与 AI 对话试图寻找多元线性回归的创新突破点

我应该如何在多元线性回归中做出创新

已深度思考 (用时 21 秒) >

非常好的问题！在数学建模竞赛中，使用多元线性回归这类基础模型时，做出创新是脱颖而出、争取高奖的关键。单纯地跑一个回归、给出系数和p值远远不够。

以下是一些在多元线性回归框架内进行创新的策略，您可以根据数据和问题的特点选择组合使用：

1. 模型设定创新：超越线性与独立

这是最核心、最有效的创新点。质疑“线性”和“相加”这两个基本假设。

- **引入非线性关系：**
 - **多项式项：**如果散点图提示可能存在曲线关系，可以加入平方项、立方项。

$$Y \sim BMI + BMI^2 + Gestational_Week$$

 - **创新解释：**这可以模拟“边际效应递减”或“加速效应”。例如，可能BMI在较低时对Y浓度影响不大，但超过某个阈值（如30）后，负面影响急剧增大。你可以通过求导找到这个“临界点”。
 - **分段回归（阈值效应）：**假设临床已知BMI=35是一个重要门槛。你可以建立两个模型：

$$Y \sim Gestational_Week + BMI \quad (\text{对于 } BMI < 35 \text{ 的样本})$$
$$Y \sim Gestational_Week + BMI \quad (\text{对于 } BMI \geq 35 \text{ 的样本})$$

 - **创新解释：**比较两个模型的系数（尤其是截距项和BMI的系数）是否有显著差异。这能有力地证明对于不同BMI人群，孕周与Y浓度的关系模式根本不同，这直接引向了问题2和3的分组需求。
- **引入交互效应：**
 - **添加交互项：**

$$Y \sim Gestational_Week + BMI + Gestational_Week * BMI$$
 - **创新解释：**这检验了“孕周和BMI是否存在联合效应”。例如，交互项显著为负意味着：对于高BMI孕妇，每增加一周孕周所带来的Y浓度提升幅度，要小于低BMI孕妇。这是一个非常深刻且有临床意义的

在探索建模过程中，此次 Deepseek 提到了分段，这启发了我们第二问以及第三问的建模过程，我们通过风险程度的分段构建了风险概率模型以及风险程度的量化。

4. 模型构建阶段，询问 AI 有哪些可能与 Y 染色体浓度相关的指标



在思考与 Y 染色体浓度有关的因素时，我们询问了 Deepseek，期望能得到一些方向。对于 AI 的解答，我们选择了纳入考虑并检验相关性，而后继续探索新的可能的有关因素。

5. 第四问构建模型阶段，让 AI 解释 GC 含量的意义以及与 Y 染色体浓度的关系

本题中的GC浓度和Y染色体浓度有关吗

已深度思考 (用时 15 秒) >

这是一个非常专业且重要的问题。简短的回答是：

有关系，但这种关系主要是技术性的、间接的，而非生物性的、直接的。在建立预测Y染色体浓度的模型时，通常不将GC含量作为核心自变量。

下面进行详细解释：

1. GC含量是什么？

- **定义：** GC含量是指测序序列中鸟嘌呤（G）和胞嘧啶（C）两种碱基所占的比例。它是评估测序数据质量的一个重要技术指标。
- **正常范围：** 如题目附录所述，正常GC含量范围为40%~60%。过高、过低或分布异常都意味着测序质量可能存在问题。

2. GC含量如何“间接”影响Y染色体浓度的计算？

Y染色体浓度（即胎儿游离DNA比例）并非直接测量出来，而是通过高通量测序获得数百万条DNA读段（Reads），再经过一套复杂的生物信息学流程计算出来的。

这个流程的关键步骤包括：

1. 将测序得到的读段与人类参考基因组进行比对。
2. 统计比对到各个染色体上的读段数量。

在解决第四问的过程中，要考虑的变量过多，且变量的具体含义我们并不是很明白，于是求助了 AI，并将其回答纳入考虑范围。

6. 第四问模型构建阶段，使用 AI 寻找 GC 含量的相关文献

GC含量的正常范围是多少，给出参考文献

已深度思考 (用时 19 秒) >

理解您想确认GC含量的正常范围。GC含量 (Guanine-Cytosine Content) 通常指DNA分子中鸟嘌呤 (G) 和胞嘧啶 (C) 两种碱基所占的比例²。这个值一般用百分比表示。

以下是不同场景下GC含量的一般范围和建议：

应用场景	一般建议范围	说明与注意事项
PCR引物 ⁹	40% - 60% (理想范围约45%-55%)	上下游引物的GC含量不应有显著差异，以保
测序引物/杂交探针 ²	至少50%	较高的GC含量有助于提高结合稳定性和特异
生物基因组		
~ 真核生物 ⁷	约 30% - 50%	不同物种差异较大，人类基因组的GC含量平
~ 原核生物 (细菌) ⁷	约 20% - 75%	GC含量是细菌分类的参考指标之一，通常认

 GC含量为何重要

GC含量之所以重要，主要是因为G和C之间通过三个氢键连接，而A和T之间只有两个氢键。因此：

- **稳定性影响：**GC含量越高，DNA双链结构就越稳定²，其解链温度 (Tm值) 也越高。
- **技术操作考量：**在设计PCR引物⁹、测序引物或杂交探针时²，GC含量会影响其与模板结合的稳定性 和特异性。因此，将其控制在合适的范围内 (通常40%-60%) 有助于成功^{2 9}。
- **基因组特征：**不同生物甚至同一生物基因组的不同区域，GC含量都可能存在差异⁷。这些差异有时与 基因密度、染色质状态乃至环境适应有关⁷。

在第四问的构建过程中，我们不确定该如何理解并使用 GC 含量，可我们无法将 AI 的回答与题目所给数据进行有效地关联起来。