数据预处理

**核心数据编码**

对于已有附件给出的数据，为便于后续更加系统性的建模分析，优先进行一系列的数据清洗，团队主要处理了以下几个方面：

1.检测孕周：统一转换为“孕期天数”统一单位，以便与检测日期进行对齐分析、和后续。

2.怀孕孕次：将怀孕次数≥3的样本统一归为3次，以减少极端值影响；

3.受孕方式：自然受孕编码为1，IUI（人工授精）编码为2，IVF（试管婴儿）编码为3。

4.胎儿健康状态：健康编码为1，异常编码为2。

5.将末次月经同每一次检测的日期进行关联，日期差值命名为“检测日期同最后一次月经的时间差”替换掉末次月经特征，随后将检测日期特征移除。

操作代码详见...

**缺失值填补**

经上数据清洗后，发现部分样本未记录末次月经（LMP）日期（男胎检测数据中编号为A108、A139、A159），但记录了检测日期及孕周，故根据已有完整信息完全可采用确定性反向推算补全缺失值，具体求解即：



|  |  |  |  |
| --- | --- | --- | --- |
| 孕妇代码 | 检测日期 | 孕期天数 | 末次月经时间 |
| A108 | 2023-10-23 | 89天 | 2023-07-26 |
| A139 | 2023-04-06 | 85天 | 2023-01-11 |
| A159 | 2023-06-15 | 95天 | 2023-03-12 |

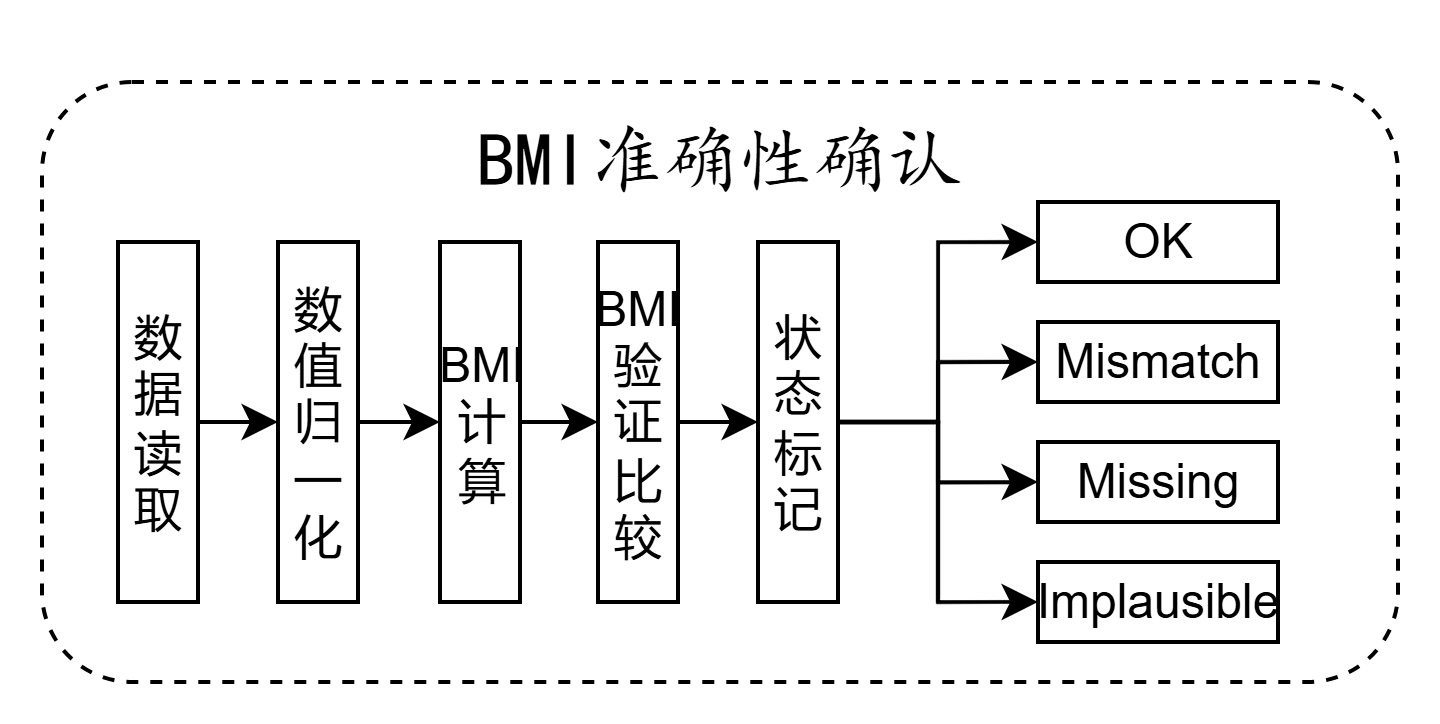
**BMI数据校验**

BMI由身高与体重计算得出，所有人群通用的计算公式为：



即某人的BMI=体重（kg）/[身高（m）]2，其中为该人的体重，体重单位为千克，为该人的身高，单位为米。

为确保数据准确性，团队建立校验流程用于验证已有孕妇BMI特征的准确性，具体流程如图所示：



*Step*1:读取原始Excel数据，解析数据特征。

*Step2*:考虑数据单位不一，为方便计算我们将身高和体重进行单位换算。

*Step3*:计算BMI并保留一位小数。

*Step4*:BMI验证比较

差值公式：



其中为自主评测差值\_1，为当前孕妇的BMI指标。

*Step5*:标记状态为：一致（OK）、不匹配（Mismatch）、缺失（Missing）或异常（Implausible）。



此处的团队设置为足够反应数据是否异常的0.1，对于可能存在的异常BMI值，设置警戒值:



当测量值超出最大警戒值或小于最小警戒值，即：



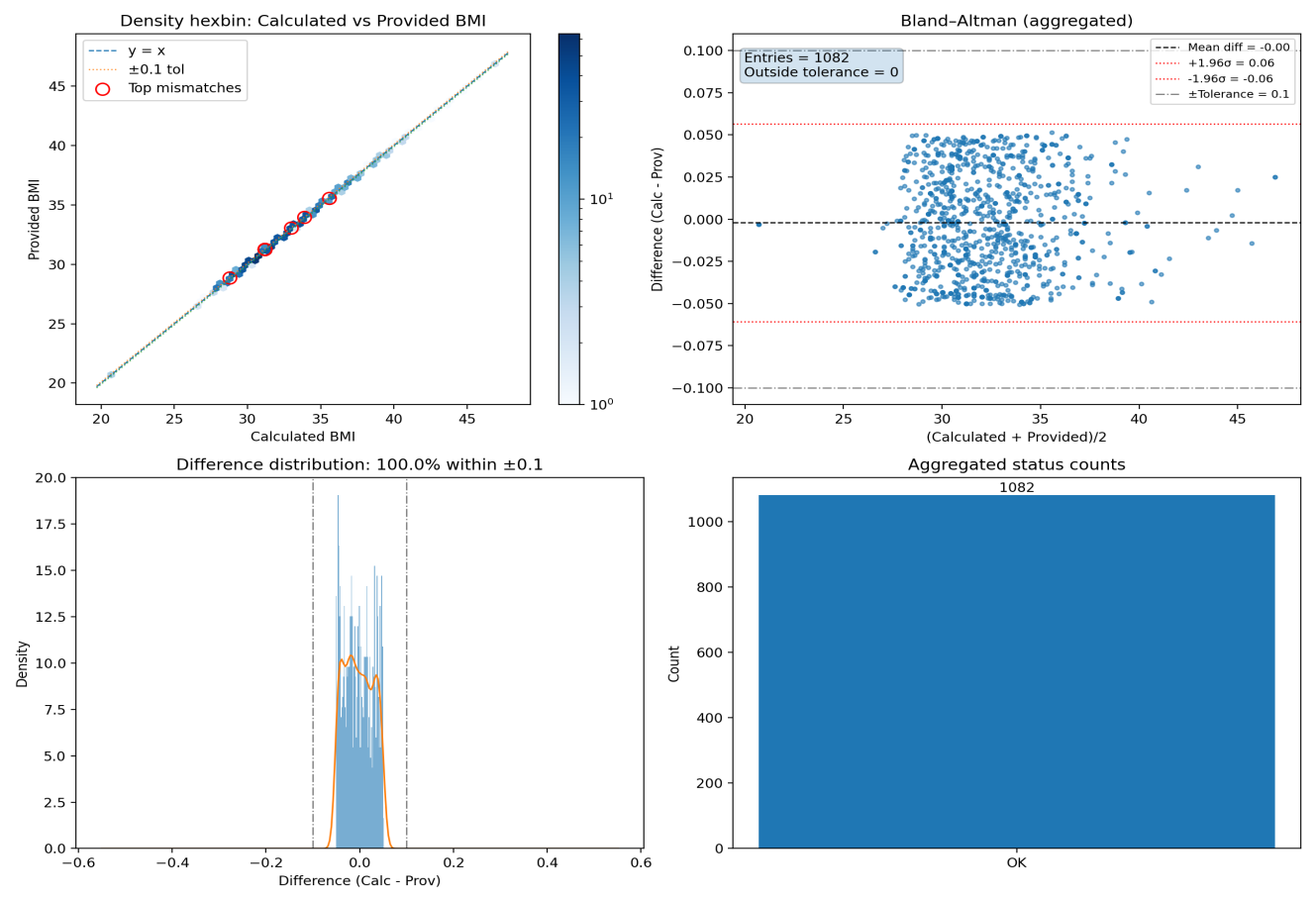
将提供预警标记。

下图为BMI数据校验模型可视化结果，通过展示登记值与计算值的密度六边形，在同张图上用点密度显示计算得到的BMI (Calculated) 与表中提供的孕妇BMI (Provided)的整体对应关系，快速看出一致性与集中趋势(比单纯散点对大量点更清晰) 。

通过Bland-Altman一致性图可视化，检验计算值与表中值的一致性， 判断差值是否随BMI大小存在系统性偏差，并提供统计一致性区间

通过差值的直方图及核密度估计定量显示差值(Calculated - Provided)的分布形态，便于判断误差是否近似正态是否有偏

校验状态计数条形图把所有测量对按简单分类做最终汇总可视化，便于快速量化数据质量。



如图示密度六边形图：大部分hexbin (颜色最深的区域)沿着对角线，说明登记值与计算值总体致数据质量好，如图中空心红点标记的 TOP mismatches，为绝对差值最大的若干记录仍然分布在对角线周围。

从Bland-Altman一致性图中，可以看出绝大多数散点落在 95% 一致性界限（LOA）内，表明登记系统与计算系统整体一致性良好，可见所有点均落在预容差内（±0.1）。

从差值的直方图及核密度估计，用以量化误差分布特征，可以发现密度曲线近似正态且中心位于0，表明差异呈随机、无偏分布，符合测量误差的高斯假设且数据质量理想，无长尾或极端离群值。

以校验状态计数条形图，将所有测量数对按数据按简单分类划分为 OK、Mismatch、Missing 及 Implausible 四类，条形图显示每类的绝对计数，可以看到完整的图中全部都是OK类型，表名数据质量高，未有较大误差。

通过以上模型方法联合验证了数据中孕妇BMI的准确性、合理性，系统性的完成了BMI数据校验。

**异常数据修复**

对异常值进行人工复核，必要时剔除或修正。

数据中发现A160检测日期同最后一次月经的时间差值为负数（-103），团队注意到这一错误，追溯原附件发现男胎检测数据中编号为A160的孕妇其末次月经时间存在混乱，第一次和第四次同二三次不一，团队通过确定性计算将该异常值修复，统一了其末次月经时间。

四、符号说明

|  |  |
| --- | --- |
| 符号 | 说明 |
| ， | 数字下标 |
|  | 女胎产妇编号 |
|  | 成品的调换损失 |
|  | Y染色体浓度 |
|  | X染色体浓度 |
|  | 孕妇BMI |
|  | 检测日期同最后一次月经的时间差 |
|  | 检测孕周 |
|  | 检测抽血次数 |
|  | 怀孕次数 |
|  | 生产次数 |
|  | 自主评测差值\_1 |
|  | 体重 |
|  | 总成本 |

**问题** 1 试分析胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 等指标的相关特性，给出相应的关系模型，并检验其显著性。

对于我们数据处理后的附件数据，已知其为1083行数据，34项特征的表格文件，除标号外，其余均已完成映射，团队初步采用皮尔逊相关系数进行相关性基本判断尝试，具体步骤如下：  
 *Step*1:剔除序号、孕妇代码等明显无关特征后形成1083x30拟定数据集，考虑到我们的数据集中既有连续变量又有0/1变量的情况，团队对拟定数据集进行Min-Max Scaling归一化处理统一量纲，初步进行相关性分析的同时最大程度保留原有数据特征，也便于后续建模过程中神经网络的应用。

Min-Max Scaling：



其中表示为数据中第项特征的第项数据，和分别表示第项特征中的最大值和最小值。特别地，上述归一化对于0/1变量特征不做缩放。归一化完成后所有数据都被映射至。

*Step2*:群体特征相关性测试，将特征y染色体浓度单独拎出，同其他所有群体特征统一进行皮尔逊相关系数判断（Pearson's correlation coefficient）：



其中为y染色体浓度同任意特征的皮尔逊相关系数，为样本协方差：



为样本的标准差：



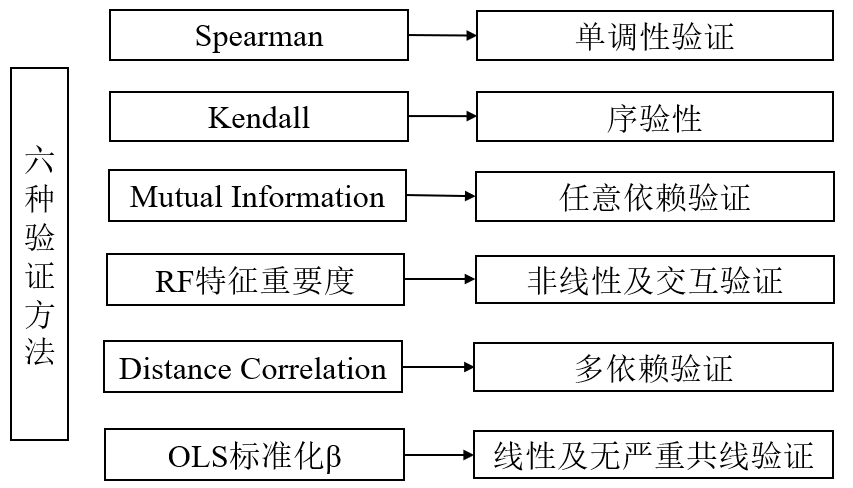
其中为当前特征中第项数据。

经计算结果如下表所示：

Pearson |r| 与 Y染色体浓度最相关 Top10

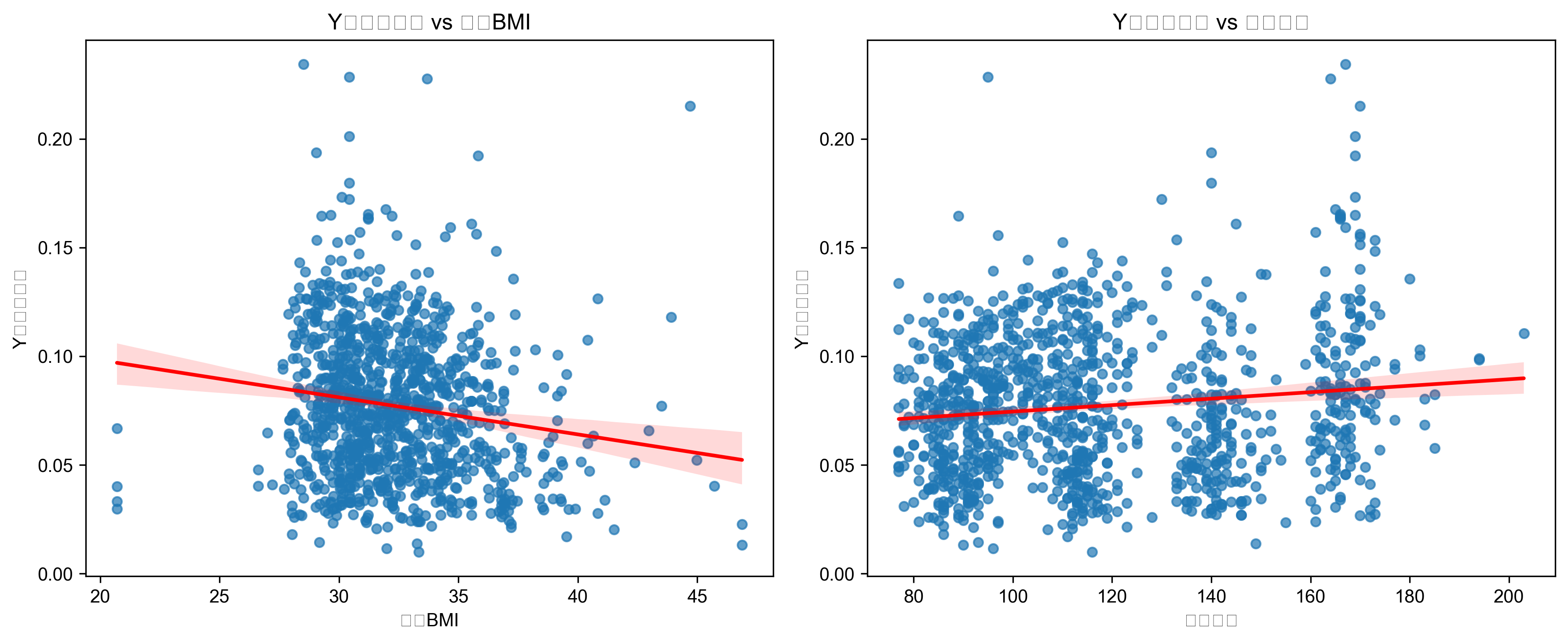
|  |  |
| --- | --- |
| Pearson | Y染色体浓度相关系数 |
| X染色体浓度 | 0.518596 |
| 检测抽血次数 | 0.329890 |
| 体重 | 0.180601 |
| 18号染色体的Z值 | 0.165558 |
| 孕妇BMI | 0.151300 |
| 检测孕周 | 0.126560 |
| 年龄 | 0.119391 |
| 原始读段数 | 0.109941 |
| 在参考基因组上比对的比例 | 0.105497 |
| 身高 | 0.104651 |

团队发现所得数据相关性近乎都极弱（详见...），这显然与材料不符，团队继续尝试进一步采取六种方式进行相关性判断研究，如下图所示：



通过数据观察（详见附件...）均无特别显著，普遍为低相关，仅有少部分呈现中度相关，这显然与材料和事实不符，团队继续深入研究探讨相关性。

*Step3*:关于个体内特征相关性联合群体分析探讨，先前对于群体综合数据分析下来各类特征同我们的目标特征均显示关联性低，这可能是由于个体间差异，在不同子群之间中特征关系反映方向或者程度不同，在合并之后相关性幅度被一定程度减弱，使得结果可能被稀释或具有误导倾向。



上图为Y浓度同孕妇BMI以及检测孕周关系曲线图。图中数据分布较为散乱，回归线较为平稳，说明了我们的群体综合关系分析存在阈值或某一限制。

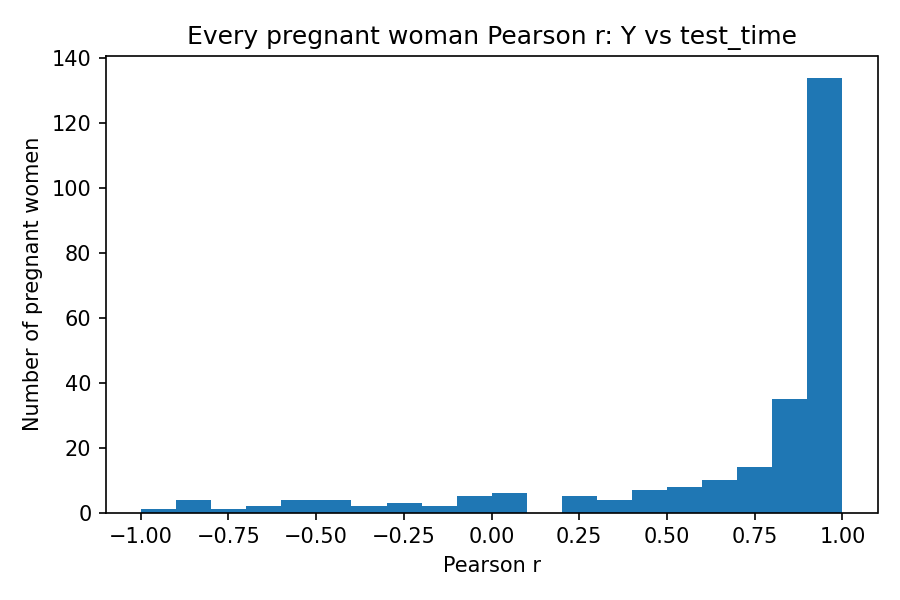
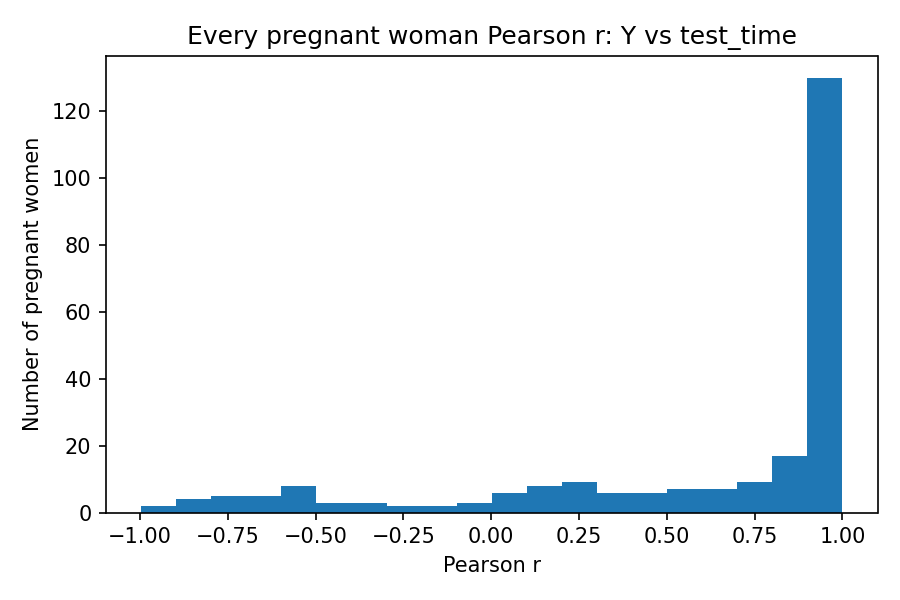
通过数据观察可视化，男胎检测数据中267名孕妇成员中有超过240最高至251名孕妇检测抽血次数大于等于三次，足够进行个体内相关研究。

|  |  |  |
| --- | --- | --- |
|  | 孕妇BMI | 检测孕周 |
| 皮尔逊相关系数平均值 | 0.6064249064886603 | 0.6943639443878687 |
| 皮尔逊相关系数中位数 | 0.9283195312094739 | 0.914595390414964 |
| 皮尔逊显著占比（p<0.05） | 0.5206611570247934 | 0.3904382470119522 |
| 斯皮尔曼秩相关平均值 | 0.6041897270047051 | 0.7031845261612758 |
| 斯皮尔曼秩相关中位数 | 0.9486832980505139 | 1 |
| 斯皮尔曼显著占比（p<0.05） | 0.5206611570247934 | 0.5776892430278885 |
| 肯德尔秩相关平均值 | 0.5827461419484785 | 0.675621912556165 |
| 肯德尔秩相关中位数 | 0.912870929175277 | 1 |
| 肯德尔显著占比（p<0.05） | 0.04132231404958678 | 0.043824701195219126 |

如上表列出所示，通过个体内特征相关性采用关联三方法模型联合进行求解，每行为“一个特征在所有孕妇个体内相关结果”的汇总统计对每一位孕妇分别计算Y染色体浓度与其他所有数值特征的、、相关性，（详见...），同时也群体层面也给出数据文件（详见...）便于对比个体内和个体间差异。

表中展示模型三大相关性均超过0.5，且随求证方法越严格相关性几乎不变，均为强相关。这表明：1.在相关性方向检测孕周越大，Y染色体浓度越高。2.孕妇BMI越高，Y染色体浓度也倾向于越高。无论是孕妇BMI还是检测孕周都存在中位数高于平均值的右偏现象，但检测孕周右偏现象程度比BMI轻，平均值和中位数都很高，表明检测孕周与Y染色体浓度之间存在的较强正向线性关系比BMI更稳定。同时，在中位数上三者几乎都是1，这是一个非常强烈的信号支持数据Y染色体浓度中存在的同孕妇BMI和检测孕周的强大单调趋势或者完美的递增关系。

除了上述表格展示的同和关系之外，还有三项特征（检测抽血次数、检测日期同最后一次月经的时间差、体重）较为相关，其余几乎不相关（模型中所有相关系数几乎均小于0.2）详见...。前两者实际意义上是检测时间的反映，后者是的反映，事实上**强有力论证了胎儿 Y 染色体浓度仅仅只与孕妇的孕周数和孕妇BMI相关。**而****与其它数据特征的相关性仍有待探究。



上图为与两关键特征和的个体特征中皮尔逊相关系数分布情况的直方图，可以看到图中直方分布大量个体集中分布在皮尔逊相关系数超过0.75的部分，但是除此之外仍然有少部分均匀分布于图中其它处。这更进一步表明了不同孕妇间个体的个性化差异非常大，但是总体上几乎可以确定与两关键特征和在大部分个体上存在极强的线性相关性。

*Step4*:关系模型的初步确认、检验和分析。

前文已经确认孕妇之间个性化差异显著，团队初步考虑建立混合效应模型，建立Y染色体浓度与两个显著相关的自变量——检测孕周和孕妇 BMI之间的关系模型，同时考虑每位孕妇（个体）有多次检测、个体差异显著的事实。

模型的初步设定为单一模型同时估计群体平均效应与个体差异即随机效应：  
****

其中，表示第次观测，表示第位孕妇。，，为群体待估效应参数。，，代表第位孕妇的随机效应（随机截距），其中。为观测误差即残差。为随机效应的协方差矩阵。

我们通过最大似然和REML估计出，随机效应协方差矩阵、残差方差 。

通过对和都随机的方式，最终固定效应已估计：



模型的随机效应和方差分量也做了估计：

但模型没有收敛：



经过参数检查报告（详见...）这是由于我们数据量中变量尺度大，而随机斜率非常小的原因，导致协方差矩阵条件数大，使得优化器难以找到稳定解。以及我们的自变量没有中心化，使得随机效应协方差估计不稳定。

*Step4*:关系模型的最终确认、检验和显著性分析。

为了解决不收敛的问题，团队依次通过三个“模型升级”来最终达到拟合模型的目的。

变量中心化，我们对自变量做全局中心化，减去样本均值，使得截距表示在自变量均值处的预测值，减少截距与斜率之间的数值相关性，改善协方差矩阵的条件数，从而提高优化器的稳定性与收敛性。





简化随机结构，优先对随机斜率，仅保留固定效应。这是因为通过先前的经验和数据（孕妇个内部相关性很高），我们发现在个体间的斜率存在异质性，而的个体斜率相比下并不显著。

数学模型对应更改为：

****优化器的不同尝试和增加迭代次数以及REML方法使用，由于有时不同优化器能跳出局部数值困难点获得收敛结果，于是团队在拟合时尝试lbggs、powell等不同优化器，增加maxiter，并在必要时引用RENK方法拟合，对于方差分量估计更稳健。

最终拟合模型，经过上述优化后，我们最后拟合且已收敛，如下表所示为模型详细输出参数：

|  |  |  |
| --- | --- | --- |
| 模型收敛状态：true | | |
| 截距（Intercept） | | 0.078413 |
| 孕周中心化值（gest\_week\_c） | | 0.000476 |
| BMI中心化值（Bmi\_c） | | -0.001193 |
| 随机效应协方差矩阵（Random effects covariance, cov\_re） | | |
|  | Group | Gest\_week\_c |
| Group | 0.000841 | 1.8989448e-06 |
| gest\_week\_c | 0.000002 | 7.415985e-08 |
| 残差方差（Residual variance, scale）：0.08828793679824699838 | | |
| 模型拟合指标 | | |
| 对数似然值（Log-Likelihood, LL） | | 2686.274689447126 |
| AIC（赤池信息量准则） | | -5198.549378894252 |
| BIC（贝叶斯信息量准则） | | -5163.643413678487 |
| 模型解释力指标 | | |
| 边际 R²（Marginal R²） | | 0.41968879873746723 |
| 条件 R²（Conditional R²） | | 0.5214619611827418 |
| 近似组内相关系数（Approximate ICC） | | 0.801765398628388 |
| 样本内预测误差 | | |
| RMSE（均方根误差） | | 0.83217746677791483 |
| MAE（平均绝对误差） | | 0.026793114661673972 |

团队模型成功拟合：  


拟合时用到的中心化常数（样本均值）如下：  


将估计值带入模型：  
**** 同时的，我们得到了随机项分布

随机效应向量：







残差方差：

以及一些重要的衍生统计量：







随机效应标准差：





随机截距与随机斜率相关约为

方差分解：



约为的总变异来自孕妇之间的系统差异（随机截距）

预测误差(样本内)





模型衡量







将“中心化模型”换回原始变量表示：

对于先前给出的中心化形式展示模型是便于数值拟合与解释的，如下：



我们可以通过展开常数项

计算（数值）：





因此合并后常数项约

于是我们得到了最终的函数模型等价形式：



这就是问题一胎儿 Y 染色体浓度与孕妇的孕周数和 BMI 相应的关系模型。

在该问题中，我们通过多模型多方法联合使用共同论证了Y 染色体浓度与其它特征的相关性，特别的论证了Y 染色体浓度与孕妇的孕周数和孕妇BMI的高度相关。通过ICC的判断：

（近似）（基于随机截距方差）：



ICC高达0.8018，这也说明约80%的变异来自孕妇之间的系统差异，混合模型提供的随机效应是必须的。

随机效应方差矩阵：  




其中非零且合理，协方差非常小但存在，这表明随机结构被数据支撑。

说明绝大多数变异来自孕妇之间，不同的孕妇个体间差异显著，在生物学解释上说明其个体间差异占主导。

我们通过一系列模型联合使用和判断，利用混合模型同时估计成功突破了群体间基线/斜率异质性把信号掩盖，导致相关性难以发现的状况。模型已收敛可信，模型展示中固定效应为0.000476且显著，这说明群体平均效应稳定存在。我们的方差分解提供的模型解释力：



固定效应已经解释了相当一部分编译，加入随机效应进一步提升解释力，模型结构的合理性得到体现。

**问题 2** 临床证明，男胎孕妇的 BMI 是影响胎儿 Y 染色体浓度的最早达标时间（即浓度达到或超 过 4%的最早时间）的主要因素。试对男胎孕妇的 BMI 进行合理分组，给出每组的 BMI 区间和最佳 NIPT 时点，使得孕妇可能的潜在风险最小，并分析检测误差对结果的影响。

A003

缺点

关键发现：显著占比极低 (BMI: 4.13%， 孕周: 4.38%)。

这与前两种方法的结果形成了巨大反差。平均值和中位数看似不低，但如此低的显著占比意味着这些相关的“置信度”非常低。

可能的原因：

样本量问题：可能每个单独的相关性分析是在小样本上进行的（例如，按不同地区、不同医院分组计算），小样本量很难达到统计显著性。

数据噪声大：Y染色体浓度的测量可能受非常多其他因素干扰（如检测技术、母亲自身遗传背景、孕妈疾病等），导致数据信噪比低，难以检测出稳定的关系。

关系不稳定：这种相关性可能只在特定条件下成立，而在整体样本中非常微弱且多变。

后面三个染色体有空缺填补可以做文章  
  
 最后可以动神经网络拟合