三、 符号说明

符号	符号说明
Y	男胎 Y 染色体浓度
Y^*	Y 的 logit 变换值
p_0	男胎达标概率阈值
F	F 统计量
B	Bootstrap 重采样轮数
D_i	Cook 距离
k	Weibull 形状参数
λ	Weibull 尺度参数
t_q	q 分位数对应的时间
β	AFT 回归系数
t_g	组内最早达标时间/最佳 NIPT 时点
$\mathbf{x}_i$	第 i 个样本的特征向量
y_i	第 i 个样本的响应/目标
γ	XGBoost 复杂度系数
δ	SMOTE 插值随机系数

四、 问题分析

4.1 问题 1 的分析

本问旨在刻画男胎样本的 Y 染色体浓度与检测孕周、BMI 及染色体非整倍体状态（Aneu）的关系，并完成模型显著性检验。首先基于题意查阅相关文献，将 Aneu 作为除孕周与 BMI 之外的关键控制变量；随后用 **logit 变换**对 Y 染色体浓度进行数据变换与刻度统一，并通过散点图与相关性初筛确认相关特征。在建模上，采用广义可加思想构造三条一维平滑项（对应孕周、BMI、Aneu），以**高斯过程回归**刻画曲线形状，并以凸权重（softmax）进行可解释集成得到总预测器。检验环节通过**近似 F 检验**评估主效应的显著性，并以 **Bootstrap 重采样**、**交叉验证**、**残差分析**检验稳健性与模型设定的合理性。在结果展示部分，输出孕周、BMI、Aneu 的平滑曲线与特征重要性表，以支撑后续按 BMI 分层选择 NIPT 时点的决策分析。

4.2 问题 2 的分析

本问要求分析男胎孕妇的 BMI 对“Y 染色体浓度首次达到 4%阈值”的时间点的影响，并将孕妇根据 BMI 分组，为各组孕妇提供最佳 NIPT 时点。建模分析如下：首先对数据样本进行多重插补以获得足够多的样本。随后，以“达标时间”为因变量、BMI 为自变量，采用**决策树回归**在均方误差准则下搜索切分点，并结合交叉验证抑制过拟合，得到稳定的**候选 BMI 分割点**。然后进一步对候选边界点进行 **K-means 聚类**，提炼出**代表性分割点**，形成 BMI 分组。分组完成后，在各组内拟合 **Weibull-AFT 模型**，推导**最佳 NITP 时点**，并通过 **Bootstrap** 法和**解析法**检测误差对结果的影响。

4.3 问题 3 的分析

在问题 2 的基础上，问题 3 在“删失建模—多重插补—分组—生存建模”的框架下引入协变量：先对删失数据进行**多重插补**，再用 Elastic-net 在 10 个候选变量中筛选与达标时间相关的**协变量**；随后以**决策树回归**获得 BMI 初始切分，并通过 **K-means** 对切分点进行稳健化聚类形成 4 组 BMI 分层；在各 BMI 组内引入所选协变量拟合 **Weibull-AFT 模型**，推导最佳 NIPT 时点；最后通过**蒙特卡洛模拟**量化测序误差对时点估计与稳定性的影响。

4.4 问题 4 的分析

本问以女胎孕妇样本为对象，将 13/18/21 号染色体非整倍体作为判定结果，综合 X 染色体及 13/18/21 染色体的 Z 值、GC 含量、读段数与相关比例、BMI 等特征建立判定模型。流程为：先按特征类型进行 **min–max 标准化**、**logit 变换**等预处理，并构造二分类指示用于重采样，同时保留多分类标签用于后续训练；针对类别不平衡，采用 **Borderline-SMOTE** 在决策边界处**过采样**以获得正常和异常样本数量平衡的训练集；随后以 XGBoost 开展多分类建模（softmax 输出、交叉熵目标与正则化），并采用十折交叉验证评估稳定性；数据划分为独立的验证测试集与按 8:2 拆分的训练子集、测试子集。最终通过混淆矩阵与 Accuracy、Recall、Precision、F1、AUC 等指标对过采样前后性能进行对比，从而形成女胎异常判定模型。

本题解决思路如下图所示：

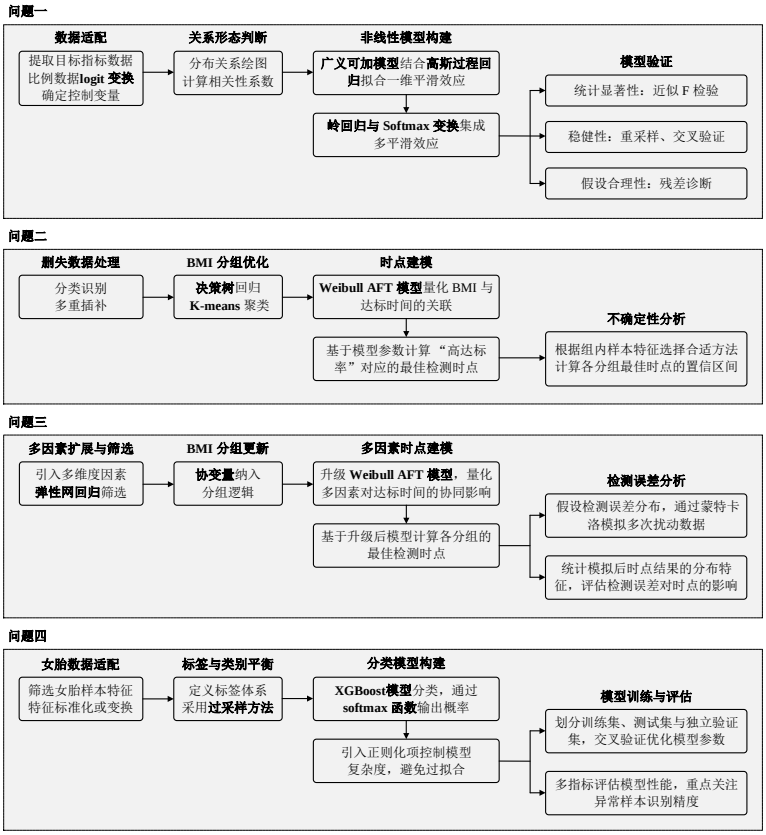

图 1 总体思路图