# NIPT 最优检测时点研究,Q3–Q4 方法说明

## Contents



# 1 符号与数据处理约定

## 1.1 符号约定

- 观测响应：$Y \in (0,1)$ 表示按比例尺度（而非百分比）的胎儿成分 proxy（男胎用 Y 染色体浓度）。
- 阈值：$y^\star = 0.04$（即 4%）为“有效检测”阈值；命中指标

$$\text{hit} = \mathbf{1}\{Y \geq y^\star\}.$$

- 协变量：$x=(t,b,a,h,w,z)$，其中

  $t$：检测孕周（周），　$b$：BMI，　$a$：年龄，
  $h$：身高，　$w$：体重，　$z$：与测序/样本质量相关的QC指标（如GC、过滤比例等）.

- 分组：按BMI将总体分为若干区间 $g\in\mathcal{G}$，每组使用组内中位数作为"代表情景"，记为 $s_g=\big(b_g^{\text{med}},a^{\text{med}},h^{\text{med}},w^{\text{med}},z^{\text{med}}\big)$.

### 1.2 数据清洗要点（与 Q1 一致）

1. 孕周解析：将诸如"11w + 6"等格式统一换算为连续周 $t\in\mathbb{R}_+$。
2. 单位统一：若大部分 $Y$ 值 $>1$，视为百分比数据并除以 100 转为比例。
3. 合理范围与质控：筛除极端/缺失；可选保留GC区间（如 $[0.3,0.7]$）；重复检测保留首检。
4. 男胎识别：以 $Y>0.005$ 为保守阈值判定男胎，进入后续分析。

## 2 Q3：两层模型构建 $P_{\text{hit}}(t\mid x)$

Q3的目标是给出随孕周变化的"达标概率"曲线

$$P_{\text{hit}}(t\mid x)=\mathbb{P}\big(Y_{\text{obs}}(t,x)\geq y^{\star}\big),$$

并在BMI分组层面，给出"最早达标时点"与"风险–命中权衡时点"，同时量化不确定性与测量误差的影响。

### 2.1 层 A（均值模型）：$\mu(x)=\mathbb{E}[Y\mid x]$

在原始比例尺度上拟合

$$\mu(x)=\mathbb{E}[Y\mid x]\approx X(x)\,\boldsymbol{\beta},$$

其中 $X(x)$ 为以孕周 $t$ 的多项式基函数（或样条）与BMI及其交互构成的设计矩阵。一个常用的可解释设定为"$t$ 的 $d$ 次多项式 + 线性BMI + 交互项 + 线性协变量":

$$X(x)=\big[1,\,t,\,t^2,\,\ldots,\,t^d,\,b,\,a,\,h,\,w,\,t{\cdot}b,\,t^2{\cdot}b,\,\ldots,\,t^d{\cdot}b\big],$$

系数 $\boldsymbol{\beta}$ 由最小二乘估计：

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\sum_i\big\{Y_i-X(x_i)\boldsymbol{\beta}\big\}^2.$$

预测时有

$$\widehat{\mu}(x)=X(x)\,\widehat{\boldsymbol{\beta}},\qquad \widehat{\mu}(x)\in[\epsilon,\,0.5-\epsilon]\,(\text{数值裁剪，防溢出}).$$

### 2.2 层 B（观测误差）：$Y_{\text{obs}} = \mu(x) + \varepsilon$

考虑测量/技术误差，设

$$Y_{\text{obs}}(x) = \mu(x) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0,\, \sigma^2(x)\big),$$

其中 $\sigma(x)$ 的估计有三种稳健方案：

1. **Global**：用整体残差标准差 $\hat{\sigma} = \mathrm{sd}\{Y - \widehat{\mu}(x)\}$。
2. **GA-Local**：对孕周 $t$ 分箱（如步长 $0.5$ 周），在每个箱内计算残差标准差并对 $t$ 做近邻插值，记为 $\widehat{\sigma}_{\text{loc}}(t)$；对小样本/单点箱进行全局回退与边界裁剪以保证数值稳定。
3. **By-QC**：若有 QC 指标 $z$，以线性形式 $\sigma(x) \approx \alpha_0 + \boldsymbol{\alpha}^{\top} z$ 拟合残差标准差与 QC 的关系（特征标准化后最小二乘）。

### 2.3 达标概率与两类“最佳时点”

由于 $Y_{\text{obs}} \sim \mathcal{N}\big(\mu(x), \sigma^2(x)\big)$，有

$$P_{\text{hit}}(x) = \mathbb{P}(Y_{\text{obs}} \geq y^{\star}) = 1 - \Phi\left(\frac{y^{\star} - \mu(x)}{\sigma(x)}\right),$$

其中 $\Phi(\cdot)$ 为标准正态分布函数。对BMI组 $g$，在“组中位情景”$s_g = (b_g^{\text{med}}, a^{\text{med}}, h^{\text{med}}, w^{\text{med}}, z^{\text{med}})$ 下定义随孕周的曲线

$$P_{\text{hit}}(t \mid s_g) = 1 - \Phi\left(\frac{y^{\star} - \mu(t, s_g)}{\sigma(t, s_g)}\right).$$

**最早达标时点（阈值法）** 给定目标达标概率阈值 $\tau$（如 $0.90$），定义

$$t_g^{\text{target}} = \inf\big\{\, t :\ P_{\text{hit}}(t \mid s_g) \geq \tau \,\big\}.$$

**风险–命中权衡时点** 题面风险分档定义为

$$\text{risk\_level}(t) = \begin{cases} 1, & t \leq 12, \\ 2, & 13 \leq t \leq 27, \\ 3, & t \geq 28. \end{cases}$$

给定权衡参数 $\lambda > 0$，定义组别 $g$ 的目标函数

$$J_g(t; \lambda) = \text{risk\_level}(t) + \lambda\big(1 - P_{\text{hit}}(t \mid s_g)\big),$$

从而

$$t_g^{\text{risk}} = \arg\min_{t} J_g(t; \lambda).$$

### 2.4 不确定性评估与误差敏感性

**Bootstrap** 置信区间　对每个 BMI 组 $g$，重复 $B$ 次组内重采样，每次重拟合 $\mu, \sigma$ 并重算 $t_g^{\text{target}}$ 与 $t_g^{\text{risk}}$：

$$\left\{t_g^{\text{target},(b)},\ t_g^{\text{risk},(b)}\right\}_{b=1}^{B}.$$

以百分位数法给出 95% CI（或报告均值 $\pm$ 中位绝对偏差）。

$\sigma$ 敏感性　对 $\sigma$ 施加比例扰动 $s \in \{0.8, 1.0, 1.2\}$，重算 $P_{\text{hit}}$、$t_g^{\text{target}}$、$t_g^{\text{risk}}$，以量化测量误差对建议时点的影响。

## 3 Q4：检测与复检策略优化

Q4 在 Q3 的 $P_{\text{hit}}(t \mid x)$ 基础上，联合考虑“检测成本/复检成本/拖延惩罚/未达标惩罚/孕周风险”，求解首检时点与（可选）复检时点的最优策略。

### 3.1 单次检测（基线）

仅允许一次抽血，候选区间 $t \in [t_{\min}, t_{\max}]$。给定权重 $c_1$（单次成本），$\lambda$（未达标惩罚），$\kappa$（拖延惩罚/周），定义

$$J_1(t) = \underbrace{\text{risk\_level}(t)}_{\text{孕周风险}} + \lambda\big(1 - P_{\text{hit}}(t \mid x)\big) + c_1 + \kappa\,(t - t_{\min}),$$

取 $t_1^\star = \arg\min_{t \in [t_{\min}, t_{\max}]} J_1(t)$.

### 3.2 两次检测（允许复检）

策略为 $(t_1, \Delta)$，其中二检时点 $t_2 = t_1 + \Delta$。记

$$P_1 = P_{\text{hit}}(t_1 \mid x), \qquad P_2 = P_{\text{hit}}(t_2 \mid x).$$

考虑两次命中概率的相关性，令 $\alpha \in [0, 1]$，并定义

$$P_2' = \alpha P_2 + (1 - \alpha) P_1,$$

则整体成功概率

$$P_{\text{succ}} = P_1 + (1 - P_1) P_2'.$$

期望结果孕周（若二检仍未达标，也以 $t_2$ 结束流程）为

$$\mathbb{E}[T_{\text{res}}] = P_1\, t_1 + (1 - P_1)\, t_2,$$

期望孕周风险

$$\mathbb{E}[\text{risk}] = P_1\, \text{risk\_level}(t_1) + (1 - P_1)\, \text{risk\_level}(t_2),$$

期望抽血次数 $\mathbb{E}[\text{draws}] = 1 + (1 - P_1)$。给定 $c_1$（首检成本）、$c_r$（复检成本）、$\lambda$（未达标惩罚）、$\kappa$（延迟惩罚），两步策略的目标函数为

$$J_2(t_1, \Delta) = c_1 + (1 - P_1) c_r + \mathbb{E}[\text{risk}] + \lambda\,(1 - P_{\text{succ}}) + \kappa\,\big(\mathbb{E}[T_{\text{res}}] - t_{\min}\big),$$

在网格 $t_1 \in [t_{\min}, t_{\max}]$, $\Delta \in \mathcal{D}$ 上搜索最小值：

$$(t_1^\star, \Delta^\star) = \arg\min_{t_1,\Delta} J_2(t_1, \Delta), \qquad t_2^\star = t_1^\star + \Delta^\star.$$

### 3.3 输出与解释

对每个 BMI 组 $g$（或个体 $x$），报告

- 单次策略：$t_1^\star$、成功率 $P_{\text{hit}}(t_1^\star \mid x)$、目标值 $J_1(t_1^\star)$；
- 两次策略：$(t_1^\star, \Delta^\star, t_2^\star)$、$P_{\text{succ}}$、$\mathbb{E}[\text{draws}]$、$\mathbb{E}[T_{\text{res}}]$、$J_2(t_1^\star, \Delta^\star)$；
- 敏感性：对 $(\lambda, \kappa, c_1, c_r, \alpha)$ 及命中阈值 $\tau$ 的稳健性比较；
- 可视化：$P_{\text{hit}}(t)$ 曲线（含推荐竖线）、$J_2(t_1, \Delta)$ 热力图等。

## 4 实现与数值细节

### 4.1 设计矩阵与数值稳定

- 多项式次数 $d$ 一般取 2–3 即可；如样本量充足可用样条替代；
- 预测时对 $\widehat{\mu}(x)$ 做区间裁剪（如 $[10^{-4}, 0.499]$）；
- **GA-Local** 的 $\widehat{\sigma}(t)$ 用分箱（步长 0.5 week）+ 邻近插值；若箱内样本极少导致方差不可估，则回退全局 $\hat{\sigma}$ 并裁剪到 $[0.005, 0.06]$。

### 4.2 区组与代表情景

组别 $g$ 的 $P_{\text{hit}}(t \mid s_g)$ 用组内 BMI 中位数、总体（或组内）年龄/身高/体重/ QC 的中位数作为代表情景，以减少维度、提升可解释性。

### 4.3 Bootstrap 与区间估计

每次重采样在组内进行（对个体索引有放回抽样），重建 $\widehat{\mu}, \widehat{\sigma}$ 并重算时点，自然反映了建模不确定性与组间异质性。CI 可用百分位数或 BCa。

### 4.4 两次策略的相关性处理

$\alpha = 1$ 表示两次命中“近似独立”; $\alpha = 0$ 则两次强相关（第二次命中概率退化为 $P_1$）。在不知道真实依赖结构时，建议报告若干 $\alpha$ 的敏感性结果以形成“策略区间”。

## 5 小结

- Q3 用“均值回归 + 误差结构”的两层框架，将“胎儿成分阈值达标”问题转化为可随孕周连续评估的概率曲线 $P_{\text{hit}}(t \mid x)$；在 BMI 分组下定义“最早达标”与“风险–命中权衡”两类时点，并用 Bootstrap 与 $\sigma$ 扰动量化不确定性与误差影响。
- Q4 在此基础上构造“一次/两次检测策略优化”，将孕周风险、未达标惩罚、抽血成本与延迟成本统一进目标函数，通过网格搜索得到 $(t_1^\star, \Delta^\star)$ 并输出成功率、期望抽血次数与期望结果孕周等指标。