

首先绘制普通模型连续变量 age 的 HR~TIME 图还有 HR~age 图

```
fit_linear <- cph(Surv(time, death) ~ age + sex, data = data)
```

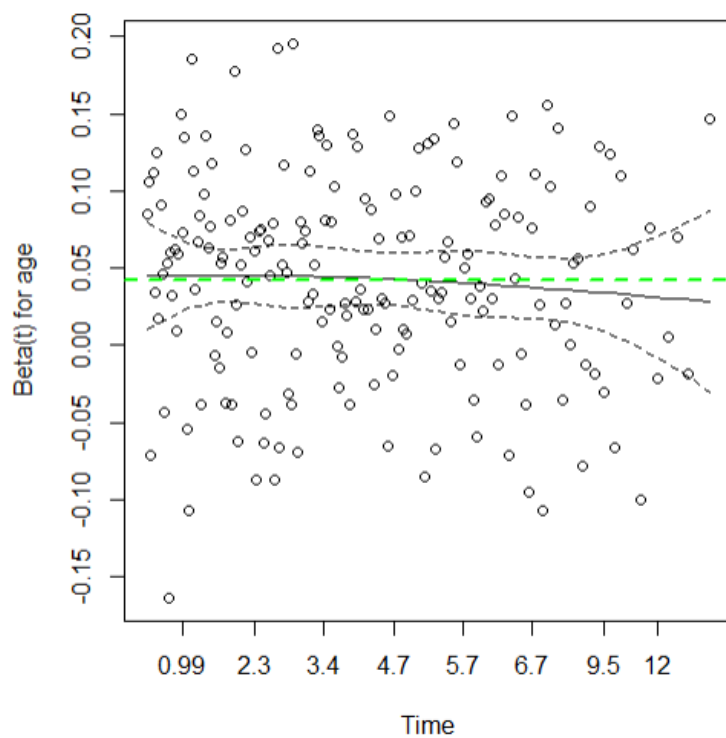
```
zp <- cox.zph(fit_linear)
```

| | chisq | df | p |
|--------|-------|----|-------|
| age | 0.597 | 1 | 0.440 |
| sex | 3.443 | 1 | 0.064 |
| GLOBAL | 3.904 | 2 | 0.142 |

```
plot(zp[1])
```

```
abline(0,0, col="red") # 0 水平线
```

```
abline(h= fit_linear $coef[1], col="green", lwd=2, lty=2) # 整体估计
```



图像进一步说明 age 满足 PH 假设检验

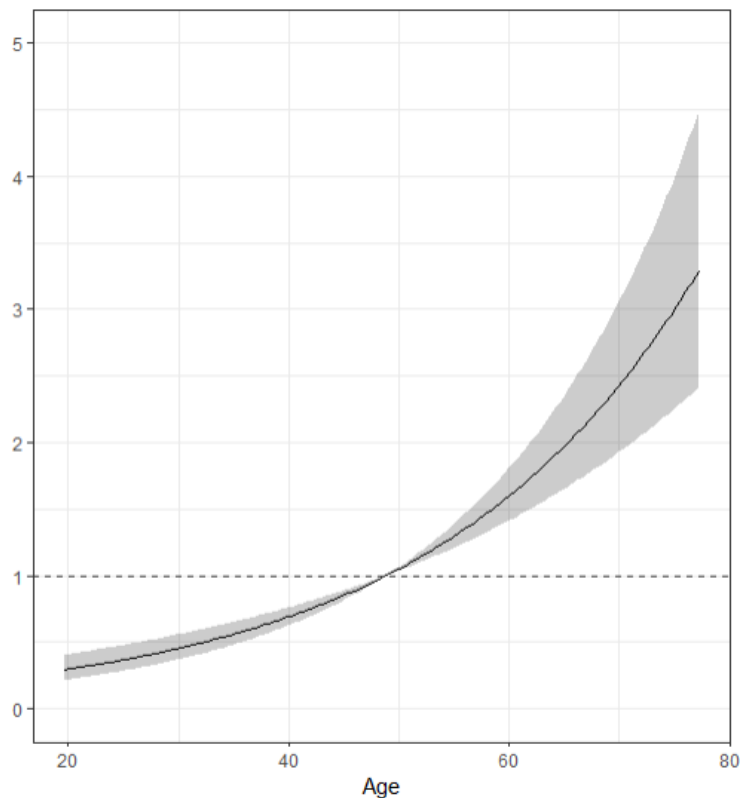
HR~age 的绘制

```
library(ggplot2)
```

```
ggplot(Predict(fit_linear, age, fun = exp, ref.zero = T))+ # fun = exp, 计算 HR 值, 不加  
这句计算概率
```

```
geom_hline(yintercept = 1, color="grey20", linetype=2)+ # HR=1 的横线
```

```
theme_bw()
```



我们可以发现随着 age 自身的变化，其 HR 值是在变化的

这时候针对连续变量的疑问点就来了，对于连续变量 age 在其满足 PH 假设检验的时候，其=每增加一岁时，其死亡增加的风险都是一样的，也就是 age 的 HR 是一定的，也就是不管你多少岁 HR 应该都是一样的，但是在 HR~age 图中，明显是随着 age 的变化，HR 在发生变化，这应该怎么解释这种看似前后矛盾的结果呢

首先绘制普通模型连续变量 age 的 HR~TIME 图还有 HR~age 图 `fit_linear <`

```
- cph(Surv(time, death) ~ age + sex, data = data) zp <-
```

```
cox.zph(fit_linear)      chisq df      p age    0.597  1 0.440 sex  
3.443  1 0.064 GLOBAL 3.904  2 0.142 plot(zp[1]) abline(0,0, col="red")
```

```
# 0 水平线 abline(h= fit_linear $coef[1], col="green", lwd=2, lty=2) # 整体
```

估计 图像进一步说明 age 满足 PH 假设检验 也就是随着生存时间的变化，

age 的 HR 一直是不变的 但是 HR~age 的绘制 `library(ggplot2)`

`ggplot(Predict(fit_linear, age, fun = exp, ref.zero = T)) + # fun = exp, 计算`

`HR 值, 不加这句计算概率 geom_hline(yintercept = 1,`

`color="grey20", linetype=2) + # HR=1 的横线 theme_bw()` 我们可以发现

随着 age 自身的变化, 其 HR 值是在变化的 这时候针对连续变量的疑问点就

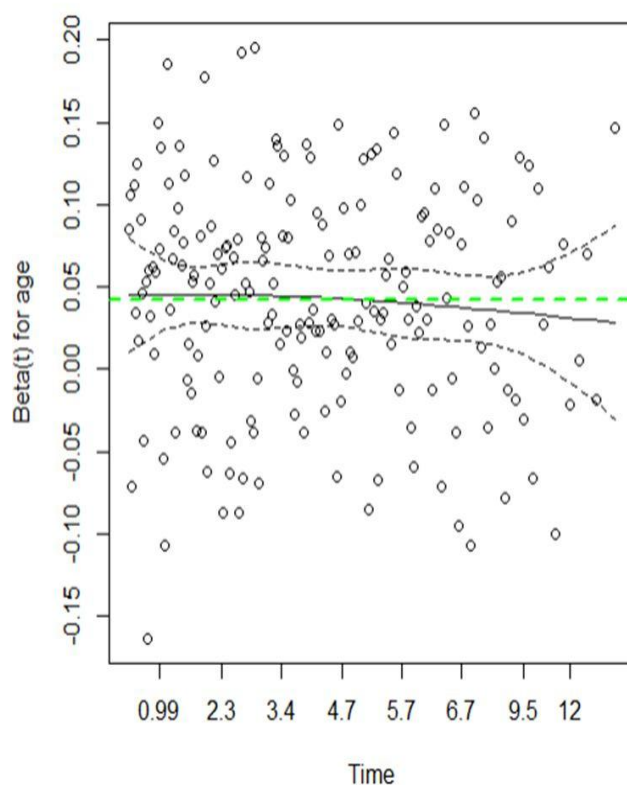
来了, 对于连续变量 age 在其满足 PH 假设检验的时候, 其每增加一岁时, 其

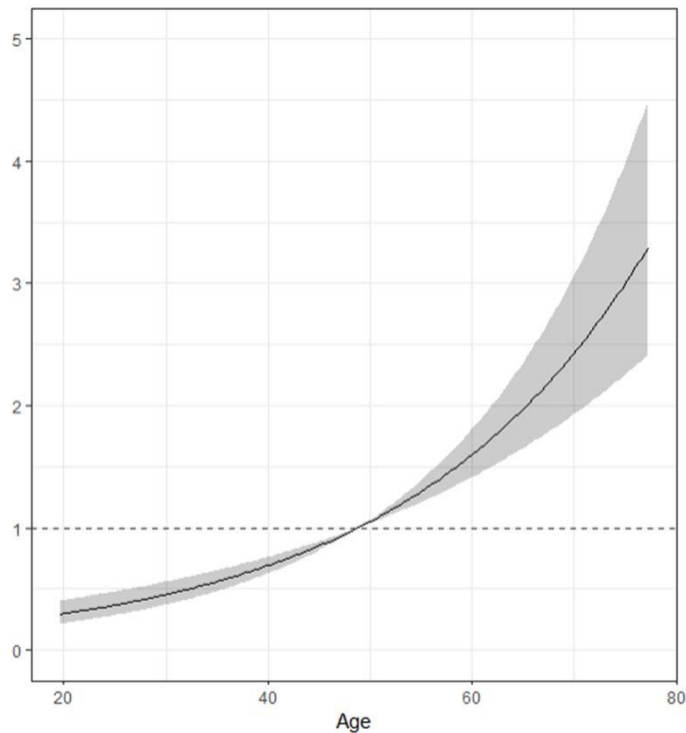
死亡增加的风险都是一样的, 也就是 age 的 HR 是一定的, 也就是不管你多少岁

HR 应该都是一样的, 但是在 HR~age 图中, 明显是随着 age 的变化, HR 在

发生变化, 这应该怎么解释这种看似前后矛盾的结果呢

用户:





首先绘制普通模型连续变量 age 的 HRTIME 图还有 HRage 图

```
fit_linear <- cph(Surv(time, death) ~ age + sex, data = data)
```

```
zp <- cox.zph(fit_linear)
```

| | chisq | df | p |
|--------|-------|----|-------|
| age | 0.597 | 1 | 0.440 |
| sex | 3.443 | 1 | 0.064 |
| GLOBAL | 3.904 | 2 | 0.142 |

```
plot(zp[1])
```

```
abline(0,0, col="red") # 0 水平线
```

```
abline(h= fit_linear $coef[1], col="green", lwd=2, lty=2) # 整体估计
```

图像进一步说明 age 满足 PH 假设检验

也就是随着生存时间的变化，age 的 HR 一直是不变的

但是

HR~age 的绘制

```
library(ggplot2)
```

```
ggplot(Predict(fit_linear, age, fun = exp, ref.zero = T))+ # fun = exp, 计算 HR 值, 不加
```

这句计算概率

```
geom_hline(yintercept = 1, color="grey20",linetype=2)+ # HR=1 的横线  
theme_bw()
```

我们可以发现随着 age 自身的变化，其 HR 值是在变化的

这时候针对连续变量的疑问点就来了，对于连续变量 age 在其满足 PH 假设检验的时候，其每增加一岁时，其死亡增加的风险都是一样的，也就是 age 的 HR 是一定的，也就是不管你多少岁 HR 应该都是一样的，但是在 HR~age 图中，明显是随着 age 的变化，HR 在发生变化，这应该怎么解释这种看似前后矛盾的结果呢

元宝：

1. 核心概念澄清

(1) PH 假设下的 HR (HR ~ Time) 加一岁和前一岁的 $\lambda(t)$ 的比值

- PH 假设检验 (cox.zph()) 验证的是：“年龄每增加 1 岁，其 HR 是否随时间变化”（即 HR 是否恒定）。
 - 你的结果中，age 的 $p=0.440$ （不显著），说明 HR 不随时间变化（满足 PH 假设）。
 - 图中 Beta(t) for age 围绕绿色水平线波动（无时间趋势），印证这一点。

(2) HR 随年龄的变化 (HR ~ Age) 现在年龄和参考年龄的 $\lambda(t)$ 的比值

- HR ~ Age 图 展示的是：“不同年龄的基线风险比（相对于参考年龄）”。
 - 即使满足 PH 假设，不同年龄的 HR 值仍可能不同，因为：
 - HR 的计算基于线性模型： $HR = \exp(\beta * \Delta age)$ 。
 - 参考点不同：图中 HR 是相对于 ref.zero（如年龄=0 岁的理论值）的比值，实际年龄越大， Δage 越大，HR 自然变化。

2. 为什么看似矛盾？

(1) 两种 HR 的含义不同

| HR 类型 | HR ~ Time (PH 假 | HR ~ Age (模型预 |
|-------|-----------------|---------------|
|-------|-----------------|---------------|

| | | |
|------|-----------------------|-------------------------------------|
| | 设) | 测) |
| 定义 | 年龄每增加 1 岁的 HR 是否随时间变化 | 不同年龄相对于参考年龄的 HR |
| 模型要求 | 要求 HR 不随时间变化 (PH 假设) | 允许 HR 随年龄值变化 (线性或非线性) |
| 数学表达 | $h(t$ | $age) = h_0(t) * \exp(\beta * age)$ |

(2) 举例说明

假设模型系数 $\beta_{age} = 0.05$ (PH 假设成立):

- **HR ~ Time:**
 - 在任何时间点, 年龄每增加 1 岁, HR 均为 $\exp(0.05) \approx 1.051$ (恒定)。
 - 例如: 50 岁 vs. 49 岁的 HR 在 $t=1$ 年和 $t=5$ 年均为 1.051。
- **HR ~ Age:**
 - 若参考年龄=40 岁, 则:
 - 50 岁的 HR = $\exp(0.05 * (50-40)) \approx 1.65$
 - 60 岁的 HR = $\exp(0.05 * (60-40)) \approx 2.72$
 - HR 随年龄绝对值变化, 但每 1 岁的增量 HR 仍为 1.051。

3. 如何统一理解?

(1) 连续变量的 HR 是“单位增量效应”

- **PH 假设** 仅要求 **单位增量 (如 1 岁) 的 HR 不随时间变化**, 但允许 **不同年龄的累积 HR 不同**。
 - 类似“利息复利”: 每年利率 (单位 HR) 固定, 但本金 (年龄基数) 越大, 总收益 (累积 HR) 越高。

(2) 可视化对比

- **HR ~ Time 图** (PH 检验):

```
plot(zp[1])  
abline(h = fit_linear$coef["age"], col="green") # 水平线表示 HR 恒定
```

- 横轴是时间，纵轴是 $\beta(t)$ (应接近水平线)。

- **HR ~ Age 图** (模型预测):

```
ggplot(Predict(fit_linear, age, fun=exp, ref.zero=TRUE)) +  
  geom_hline(yintercept=1, linetype=2) +  
  labs(y="HR (vs. age=0)")
```

- 横轴是年龄，纵轴是 HR 相对于年龄=0 的比值 (呈指数上升)。

(3) 数学关系

- 若 $h(t|age) = h_0(t) * \exp(\beta * age)$, 则:
 - **单位增量 HR**: $HR_{per_1year} = \exp(\beta)$ (恒定, PH 假设要求)。
 - **绝对年龄 HR**: $HR_{age} = \exp(\beta * age)$ (随年龄变化, 非矛盾)。

4. 总结

- **PH 假设成立**: 年龄每增加 1 岁的 HR (如 1.05) 不随时间变化。
- **HR 随年龄变化**: 不同年龄的累积 HR (如 50 岁 vs. 40 岁) 因基数不同而不同, 这是模型预期的线性效应。
- **无矛盾**: 两者分别描述 **单位增量效应** 和 **累积效应**, 本质是同一模型的不同视角。

建议: 在论文中明确区分“单位增量 HR”和“绝对年龄 HR”的解释, 避免读者混淆。

用户:

但是我通过这个代码我怎么知道我的 HR ~ Age 图中的参考年龄是多少岁呢

元宝:

在 **HR ~ Age** 图 (第 2 张图) 中, 参考年龄 (即 **HR=1** 对应的年龄) 可以通过以下方法确定:

1. 如何判断参考年龄?

(1) 默认参考年龄

- **Predict()** 函数的默认行为: 当使用 **ref.zero = TRUE** 时, 参考年龄 (HR=1 的点) 是 连续变量 **age** 的中位数 (或模型拟合时的中心化值)。
 - 例如: 若数据中 **age** 的中位数是 50 岁, 则 HR=1 对应的年龄为 50 岁。

(2) 代码验证参考年龄

通过以下代码可显式查看参考年龄:

```
# 查看模型中 age 的中心化值 (参考点)
ref_age <- mean(data$age, na.rm = TRUE) # 或 median(data$age)
print(ref_age)
```

● 输出示例:

```
[1] 50.2
```

表示参考年龄约为 50 岁 (HR=1 的点)。

(3) 从图中直接读取

- 第 2 张图的解释:
 - 图中 **水平虚线 (HR=1)** 与曲线的交点为参考年龄。
 - 若曲线在 **age=50** 时穿过 HR=1, 则参考年龄为 50 岁。
 - 若曲线整体在 HR=1 上方 (如你的图中), 说明参考年龄可能低于数据范围 (如 **age=0**)。
-

2. 为什么你的图中 HR=1 不在可见范围内？

● 可能原因：

1. `ref.zero = TRUE` 但参考年龄为 0 岁：

- 若未中心化 `age`，模型默认 `age=0` 为参考点（此时 HR=1 对应 `age=0`，但图中未显示 0 岁）。
- **解决方法：**手动指定参考年龄：

```
Predict(fit_linear, age = seq(20, 80, by=5), ref.zero = TRUE, refvalue = 50) #  
以 50 岁为参考
```

2. 数据中 `age` 的最小值 > 0：

- 若年龄范围是 20~80 岁，HR=1 的点可能在 `age=20` 左侧（不可见）。

3. 如何修改参考年龄？

(1) 方法 1：显式指定参考值

```
# 以 60 岁为参考年龄 (HR=1)  
ggplot(Predict(fit_linear, age, fun = exp, ref.zero = TRUE, refvalue = 60)) +  
  geom_hline(yintercept = 1, linetype = 2) +  
  labs(y = "HR (vs. age=60)")
```

(2) 方法 2：中心化变量

在建模前将 `age` 中心化，使参考年龄有意义（如以均值或中位数为基准）：

```
data$age_centered <- data$age - median(data$age)  
fit_linear <- cph(Surv(time, death) ~ age_centered + sex, data = data)  
# 此时 HR=1 对应中位年龄
```

4. 结合你的图片解释

- 第 2 张图 (HR ~ Age):

- 曲线从 $HR \approx 0.5$ (20 岁) 上升到 $HR \approx 4$ (80 岁), 说明:

- 若未指定 `refvalue`, 参考年龄可能是 `age=0` ($HR=1$ 在 0 岁, 图中未显示)。
 - **实际意义:** 80 岁 vs. 20 岁的 $HR \approx 4/0.5=8$ (风险高 8 倍), 但每增加 1 岁的 HR 仍恒定 (由 PH 假设保证)。
-

5. 总结

- **默认参考年龄:** 由 `mean(data$age)` 或 `median(data$age)` 决定, 可通过 `refvalue` 参数修改。
- **图中 $HR=1$ 的位置:** 若不在可见范围, 需检查是否未中心化或参考年龄设置不合理。
- **关键代码:**

```
# 显式设置参考年龄为 60 岁
Predict(fit_linear, age, ref.zero = TRUE, refvalue = 60)
```

那么到底什么时候适用样条样的 cox 回归呢

同样举例

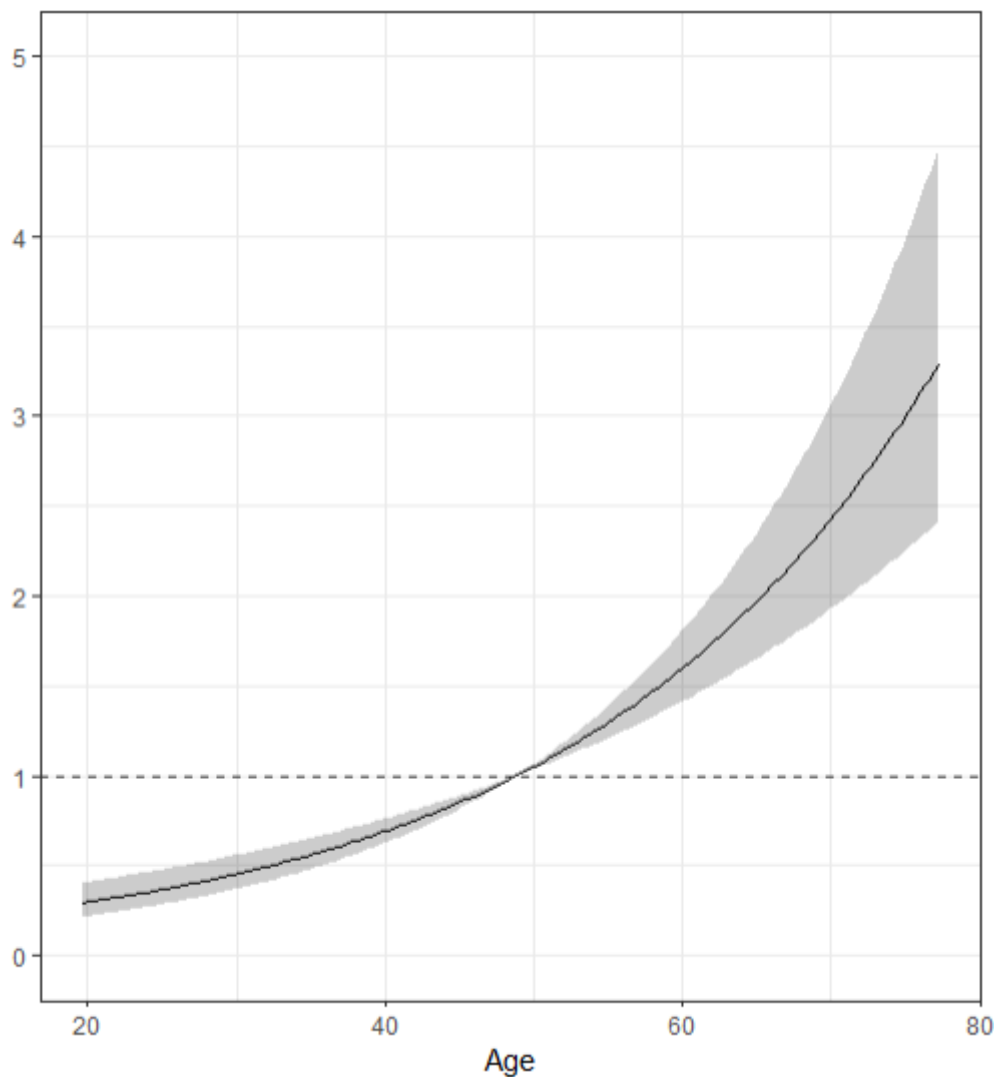
```
fit_linear <- cph(Surv(time, death) ~ age + sex, data = data)
```

```
library(ggplot2)
```

```
ggplot(Predict(fit_linear, age, fun = exp, ref.zero = T))+ # fun = exp, 计算 HR 值, 不加这句计算概率
```

```
  geom_hline(yintercept = 1, color="grey20", linetype=2)+ # HR=1 的横线
```

```
  theme_bw()
```



如果在我的简单的 cox 模型中，我的所有的连续变量的 PH 假设检验也通过，绘制出来的 HR~连续变量图也没有呈现明显的 U 型的那种明显非线性只是有一点小小的弯曲，这时候我们怎么知道对某个变量是否要进行样条化，后续对 age 进行样条处理

```
fit_spline <- cph(Surv(time, death) ~ rcs(age, 4) + sex, data = data)
```

并对后续模型进行检验

```
anova(fit_spline)
```

| Wald Statistics | | | Response: Surv(time, death) |
|-----------------|------------|------|-----------------------------|
| Factor | Chi-Square | d.f. | P |
| age | 57.75 | 3 | <.0001 |
| Nonlinear | 8.17 | 2 | 0.0168 |
| sex | 18.75 | 1 | <.0001 |

TOTAL 75.63 4 <.0001

结果显示 age 的非线性项显著

且和原本的简单模型进行比较

`lrtest(fit_linear, fit_spline)`

显示

| L.R. Chisq | d.f. | P |
|------------|------------|------------|
| 8.33954813 | 2.00000000 | 0.01545575 |

显示

若 $P < 0.05$ 提示我们的样条样的 cox 模型拟合的更好

所以我想知道的是，在一开始我们的简单模型就在我的简单的 cox 模型中，我的所有的连续变量的 PH 假设检验也通过，绘制出来的 HR~连续变量图也没有呈现明显的 U 型的那种明显非线性只是有一点小小的弯曲，这时候我们怎么知道对某个变量是否要进行样条化

1. 初步检查：何时考虑样条化？

即使满足以下条件，仍可能需要样条化：

1. **PH 假设通过** (`cox.zph p > 0.05`)：仅说明 HR 不随时间变化，但 **不保证变量与 $\log(\text{HR})$ 的关系是线性的**。
2. **HR~变量图有轻微弯曲**：提示可能存在非线性，但需进一步量化。
3. **临床原有知识点支持非线性**：如年龄对死亡风险的影响可能呈 J 型或阈值效应。——比如 65 岁以上房颤卒中风险极具增加

值得注意的是在 PH 假设检验通过的时候且曲线有点弯曲需要考虑样条样 cox

但是 PH 假设检验不通过的话，就必须对时间分层或者引入时序变量或者时序系数

思维导图如下

模型选择

检查PH假设

- $P > 0.05$ 假设检验通过 —— HR不随时间变化 —— 检查是非线性 ——
- $P < 0.05$ 假设检验不通过 —— 对时间分层或者引入时序变量系数

检查是非线性

HR~变量图有轻微弯曲：提示可能存在非线性，但需进一步量化

临床原有知识点支持非线性：如年龄对死亡风险的影响可能呈J型或阈值效应。——比如65岁以上房颤卒中风险极具增加

时序变量系数

判断是否线性

线性的 —— `fit_linear`

非线性 —— `fit_spline`

10