泊松回归

续本达

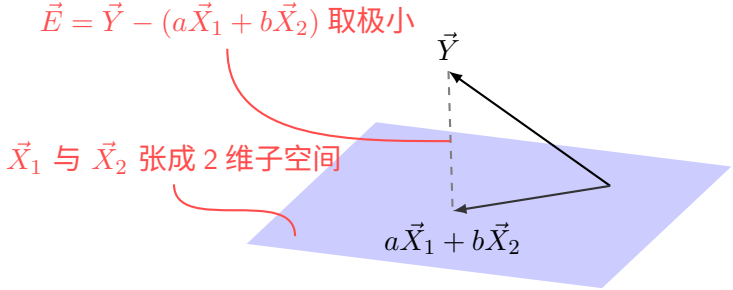
清华大学 工程物理系

2024-12-02

复习

线性回归

1. 线性回归是检验相关性的重要数理统计方法。
2. 最小二乘法是线性回归的解法，可看作线性空间的投影。



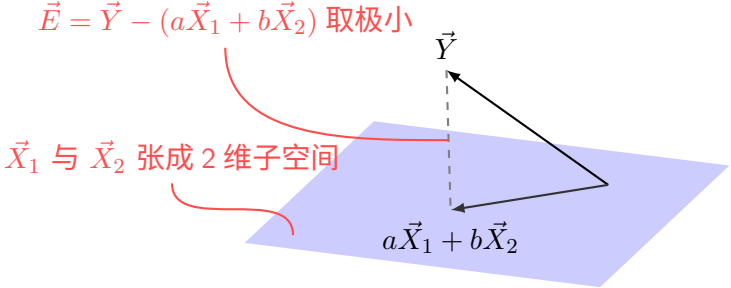
3. 假设残差来自正态总体，

$$\epsilon_i = y_i - (a + bx_i) \sim N(0, \sigma^2)$$

线性回归的结果需要经过 F 或 t 假设检验，确认回归的显著性，才能完成科学解读。

线性回归

1. 线性回归是检验相关性的重要数理统计方法。
2. 最小二乘法是线性回归的解法，可看作线性空间的投影。



3. 假设残差来自正态总体，

$$\epsilon_i = y_i - (a + bx_i) \sim N(0, \sigma^2)$$

线性回归的结果需要经过 F 或 t 假设检验，确认回归的显著性，才能完成科学解读。

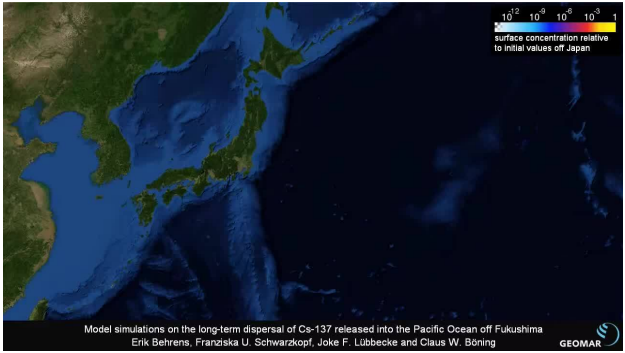
引子

推广：打破残差的正态分布假设

残差不服从正态分布时，以线性模型解决问题 $\rightarrow$ 广义线性回归。

- 泊松回归：观测量是计数
- 伽马回归：观测量是正实数
- 逻辑回归：观测量是 $[0,1]$ 区间内的实数
- 二项回归：观测量是 $0,1\cdots,N$ 的整数

例：盖革计数器的放射性计量

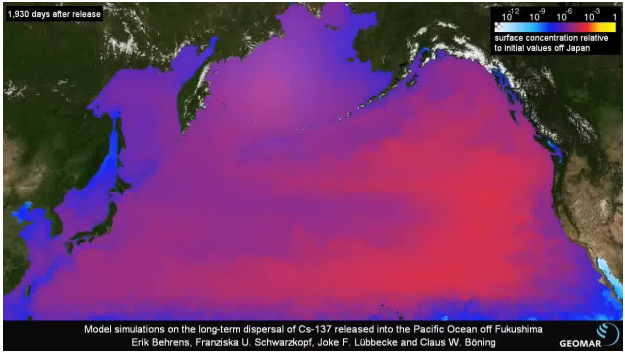


2023 年 8 月 24 日至 9 月 11 日，日本福岛第一核电站核污染水排海 7788 吨。

早在 2012 年德国海洋研究团队模拟了污水的扩散过程。

- 环境放射性，由探头计数测量。
- 计数个数反映了辐射强弱 → 样例视频。

例：盖革计数器的放射性计量



2023 年 8 月 24 日至 9 月 11 日，日本福岛第一核电站核污染水排海 7788 吨。

早在 2012 年德国海洋研究团队模拟了污水的扩散过程。

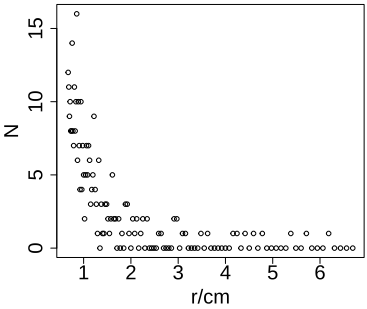
- 环境放射性，由探头计数测量。
- 计数个数反映了辐射强弱 $\rightarrow$ 样例视频。

计数问题：规律探寻

10 s 内积累的信号数 N，依赖于样品辐射强度 I 和探头距离 r。

$$N = \frac{I}{r^s}, \qquad s \approx 2$$

r 和 N 已测得，I 和 s 待求。



- 尝试转化为线性回归

$$\log N = \log I - s \log r$$

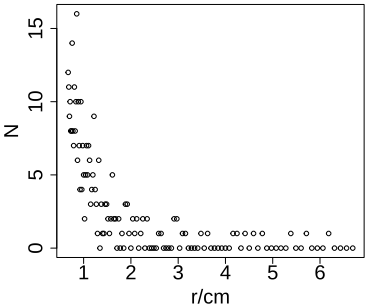
$N = 0$ 的数据点如何处理？把 $\log N$ 换成 $\log(N+1)$。

计数问题：规律探寻

10 s 内积累的信号数 N，依赖于样品辐射强度 I 和探头距离 r。

$$N = \frac{I}{r^s}, \qquad s \approx 2$$

r 和 N 已测得，I 和 s 待求。



- 尝试转化为线性回归

$$\log N = \log I - s \log r$$

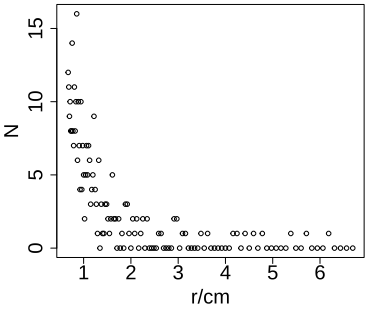
$N = 0$ 的数据点如何处理？把 $\log N$ 换成 $\log(N+1)$。

计数问题：规律探寻

10 s 内积累的信号数 N，依赖于样品辐射强度 I 和探头距离 r。

$$N = \frac{I}{r^s}, \qquad s \approx 2$$

r 和 N 已测得，I 和 s 待求。

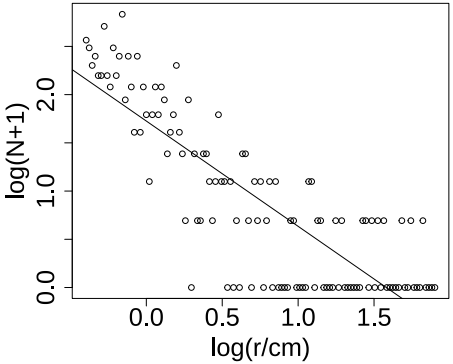


- 尝试转化为线性回归

$$\log N = \log I - s \log r$$

$N = 0$ 的数据点如何处理？把 $\log N$ 换成 $\log(N+1)$。

计数问题：线性回归



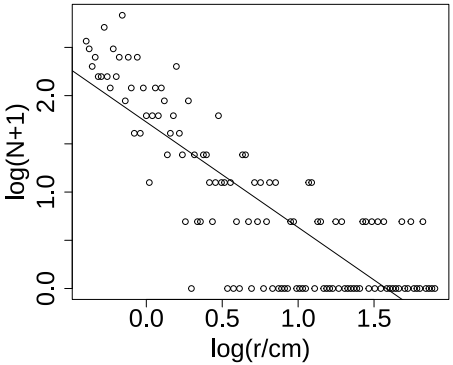
在不同距离下计数的方差不固定
计数量是离散的

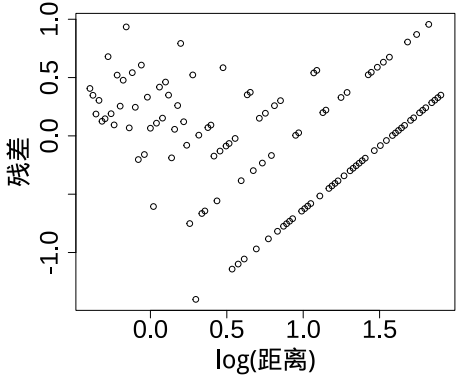
不符合线性回归的正态分布假设

$$Y_i \sim N(a + bx_i, \sigma^2)$$

方差不固定 $\xrightarrow{\text{加权线性回归}}$ $Y_i \sim N(a + bx_i, \sigma_i^2)$

计数问题：线性回归





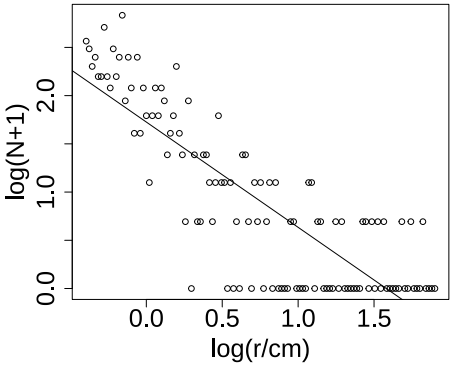
在不同距离下计数的方差不固定
计数量是离散的

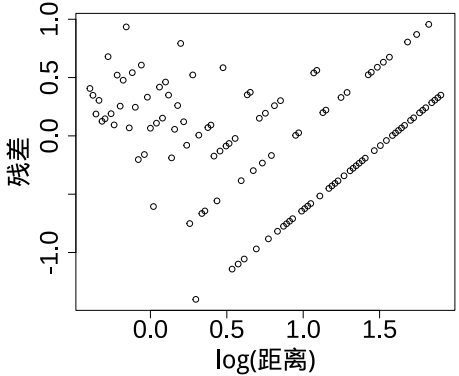
不符合线性回归的正态分布假设

$Y_i \sim N(a + bx_i, \sigma^2)$

方差不固定 $\xrightarrow{\text{加权线性回归}}$ $Y_i \sim N(a + bx_i, \sigma_i^2)$

计数问题：线性回归





在不同距离下计数的方差不固定
计数量是离散的

不符合线性回归的正态分布假设

$Y_i \sim N(a + bx_i, \sigma^2)$

方差不固定 $\xrightarrow{\text{加权线性回归}} Y_i \sim N(a + bx_i, \sigma_i^2)$

泊松回归

计数问题：把残差假设替换为泊松分布

计数问题中，残差的正态假设失效，但数据呈现依赖关系。

- 把正态分布 $N(\cdot, \sigma^2)$ 换成贴近计数问题的泊松分布 $\pi(\cdot)$

$$Y_i \sim N[\mathrm{E}(Y_i), \sigma^2], \qquad \mathrm{E}(Y_i) = a + bx_i$$

$$\text{替换为} \quad Y_i \sim \pi[\mathrm{E}(Y_i)], \qquad \log \mathrm{E}(Y_i) = a + bx_i$$

$\mathrm{E}(Y_i)$ 代表随机变量 Y_i 的期望（均值）。

定义：泊松回归

观测量 y_i 相对于预测量 $\mathrm{E}(Y_i) = e^{a+bx_i}$ 为泊松分布的回归模型，称泊松回归。

计数问题：把残差假设替换为泊松分布

计数问题中，残差的正态假设失效，但数据呈现依赖关系。

- 把正态分布 $N(\cdot, \sigma^2)$ 换成贴近计数问题的泊松分布 $\pi(\cdot)$

$$
\begin{aligned}
& Y_i \sim N[\mathrm{E}(Y_i), \sigma^2], && \mathrm{E}(Y_i) = a + bx_i \\
\text{替换为}\quad & Y_i \sim \pi[\mathrm{E}(Y_i)], && \log \mathrm{E}(Y_i) = a + bx_i
\end{aligned}
$$

$\mathrm{E}(Y_i)$ 代表随机变量 Y_i 的期望（均值）。

定义：泊松回归

观测量 y_i 相对于预测量 $\mathrm{E}(Y_i) = e^{a+bx_i}$ 为泊松分布的回归模型，称泊松回归。

计数问题：把残差假设替换为泊松分布

计数问题中，残差的正态假设失效，但数据呈现依赖关系。

- 把正态分布 $N(\cdot, \sigma^2)$ 换成贴近计数问题的泊松分布 $\pi(\cdot)$

$$
\begin{aligned}
& Y_i \sim N[\mathrm{E}(Y_i), \sigma^2], && \mathrm{E}(Y_i) = a + bx_i \\
\text{替换为} \quad & Y_i \sim \pi[\mathrm{E}(Y_i)], && \log \mathrm{E}(Y_i) = a + bx_i
\end{aligned}
$$

$\mathrm{E}(Y_i)$ 代表随机变量 Y_i 的期望（均值）。

定义：泊松回归

观测量 y_i 相对于预测量 $\mathrm{E}(Y_i) = e^{a+bx_i}$ 为泊松分布的回归模型，称泊松回归。

泊松回归模型

一组相互独立的数据 (x_i, y_i) ，其中 y_i 为计数。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i$$

- 当 x_i 确定时，a, b 决定了 Y_i 的泊松分布参数，

$$P(Y_i = y_i | a, b) = \frac{[\mathrm{E}(Y_i)]^{y_i}}{y_i!} e^{-\mathrm{E}(Y_i)}$$

$$= \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

$$P(\vec{Y} = \vec{y} | a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

- a, b 取不同的值时，观测得到的 (x_i, y_i) 的概率不同。

泊松回归模型

一组相互独立的数据 (x_i, y_i) ，其中 y_i 为计数。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i$$

- 当 x_i 确定时，a, b 决定了 Y_i 的泊松分布参数，

$$P(Y_i = y_i | a, b) = \frac{[\mathrm{E}(Y_i)]^{y_i}}{y_i!} e^{-\mathrm{E}(Y_i)}$$

$$= \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

$$P(\vec{Y} = \vec{y} | a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

- a, b 取不同的值时，观测得到的 (x_i, y_i) 的概率不同。

泊松回归模型

一组相互独立的数据 (x_i, y_i) ，其中 y_i 为计数。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i$$

- 当 x_i 确定时，a, b 决定了 Y_i 的泊松分布参数，

$$\begin{aligned} P(Y_i = y_i | a, b) &= \frac{[\mathrm{E}(Y_i)]^{y_i}}{y_i!} e^{-\mathrm{E}(Y_i)} \\ &= \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)} \end{aligned}$$

$$P(\vec{Y} = \vec{y} | a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

- a, b 取不同的值时，观测得到的 (x_i, y_i) 的概率不同。

泊松回归模型

一组相互独立的数据 (x_i, y_i) ，其中 y_i 为计数。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i$$

- 当 x_i 确定时，a, b 决定了 Y_i 的泊松分布参数，

$$\begin{aligned} P(Y_i = y_i | a, b) &= \frac{[\mathrm{E}(Y_i)]^{y_i}}{y_i!} e^{-\mathrm{E}(Y_i)} \\ &= \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)} \\ P(\vec{Y} = \vec{y} | a, b) &= \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)} \end{aligned}$$

- a, b 取不同的值时，观测得到的 (x_i, y_i) 的概率不同。

泊松回归模型

一组相互独立的数据 (x_i, y_i) ，其中 y_i 为计数。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i$$

- 当 x_i 确定时，a, b 决定了 Y_i 的泊松分布参数，

$$\begin{aligned} P(Y_i = y_i | a, b) &= \frac{[\mathrm{E}(Y_i)]^{y_i}}{y_i!} e^{-\mathrm{E}(Y_i)} \\ &= \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)} \\ P(\vec{Y} = \vec{y} | a, b) &= \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)} \end{aligned}$$

- a, b 取不同的值时，观测得到的 (x_i, y_i) 的概率不同。

对数似然距离

最大似然估计

定义复习：似然函数

观测发生的概率关于模型参数的函数称为似然函数。

- 已经测量 x_i, y_i 时，

$$P(\vec{Y} = \vec{y}|a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

是 a, b 的函数，是模型和观测值的似然函数 $\mathcal{L}(a, b)$

最大似然估计

即找到对应使 (x_i, y_i) 概率最大的一组 $\hat{a}, \hat{b}$ 为解。取 $\log$ 方便计算

$$\hat{a}, \hat{b} = \arg\max_{a,b} \log \mathcal{L}(a, b) = \arg\max_{a,b} \sum_i y_i(a + bx_i) - e^{(a+bx_i)}$$

最大似然估计

定义复习：似然函数

观测发生的概率关于模型参数的函数称为似然函数。

- 已经测量 x_i, y_i 时，

$$P(\vec{Y} = \vec{y}|a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

是 a, b 的函数，是模型和观测值的似然函数 $\mathcal{L}(a, b)$

最大似然估计

即找到对应使 (x_i, y_i) 概率最大的一组 $\hat{a}, \hat{b}$ 为解。取 log 方便计算

$$\hat{a}, \hat{b} = \arg\max_{a,b} \log \mathcal{L}(a, b) = \arg\max_{a,b} \sum_i y_i(a + bx_i) - e^{(a+bx_i)}$$

最大似然估计

定义复习：似然函数

观测发生的概率关于模型参数的函数称为似然函数。

- 已经测量 x_i, y_i 时，

$$P(\vec{Y} = \vec{y}|a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

是 a, b 的函数，是模型和观测值的似然函数 $\mathcal{L}(a, b)$

最大似然估计

即找到对应使 (x_i, y_i) 概率最大的一组 $\hat{a}, \hat{b}$ 为解。取 log 方便计算

$$\hat{a}, \hat{b} = \arg\max_{a,b} \log \mathcal{L}(a, b) = \arg\max_{a,b} \sum_i y_i(a + bx_i) - e^{(a+bx_i)}$$

最大似然估计

定义复习：似然函数

观测发生的概率关于模型参数的函数称为似然函数。

- 已经测量 x_i, y_i 时，

$$P(\vec{Y} = \vec{y}|a, b) = \prod_i \frac{[\exp(a + bx_i)]^{y_i}}{y_i!} e^{-\exp(a+bx_i)}$$

是 a, b 的函数，是模型和观测值的似然函数 $\mathcal{L}(a, b)$

最大似然估计

即找到对应使 (x_i, y_i) 概率最大的一组 $\hat{a}, \hat{b}$ 为解。取 log 方便计算

$$\hat{a}, \hat{b} = \arg\max_{a,b} \log \mathcal{L}(a, b) = \arg\max_{a,b} \sum_i y_i(a + bx_i) - e^{(a+bx_i)}$$

对数似然距离

$$\log \mathcal{L}[\mathrm{E}(\vec{Y})] = \sum_i y_i \log[\mathrm{E}(Y_i)] - \mathrm{E}(Y_i), \quad \frac{\partial \log \mathcal{L}[\mathrm{E}(\vec{Y})]}{\partial [\mathrm{E}(Y_i)]} = \frac{y_i}{\mathrm{E}(Y_i)} - 1$$

若 $\mathrm{E}(\vec{Y})$ 不受 $e^{(a+bx_i)}$ 约束，$\mathcal{L}[\mathrm{E}(\vec{Y})]$ 在 $\mathrm{E}(\vec{Y}) = \vec{y}$ 时取最大。

定义预测值 $\mathrm{E}(\vec{Y})$ 到观测值 $\vec{y}$ 的“距离” $\mathcal{D}$

- 令其越小代表两者符合得越好，取 0 时代表两者“相等”。

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = -2\left[\log \underbrace{\mathcal{L}[\mathrm{E}(\vec{Y}) = \vec{y}]}_{\text{似然函数的最大取值}} - \log \mathcal{L}[\mathrm{E}(\vec{Y})]\right]$$

$$= -2\left[\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right]$$

对数似然距离

$$\log \mathcal{L}[\mathrm{E}(\vec{Y})] = \sum_i y_i \log[\mathrm{E}(Y_i)] - \mathrm{E}(Y_i), \quad \frac{\partial \log \mathcal{L}[\mathrm{E}(\vec{Y})]}{\partial [\mathrm{E}(Y_i)]} = \frac{y_i}{\mathrm{E}(Y_i)} - 1$$

若 $\mathrm{E}(\vec{Y})$ 不受 $e^{(a+bx_i)}$ 约束，$\mathcal{L}[\mathrm{E}(\vec{Y})]$ 在 $\mathrm{E}(\vec{Y}) = \vec{y}$ 时取最大。

定义预测值 $\mathrm{E}(\vec{Y})$ 到观测值 $\vec{y}$ 的“距离” $\mathcal{D}$

- 令其越小代表两者符合得越好，取 0 时代表两者“相等”。

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = -2\left[\log \underbrace{\mathcal{L}[\mathrm{E}(\vec{Y}) = \vec{y}]}_{\text{似然函数的最大取值}} - \log \mathcal{L}[\mathrm{E}(\vec{Y})]\right]$$

$$= -2\left[\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right]$$

对数似然距离

$$\log \mathcal{L}[\mathrm{E}(\vec{Y})] = \sum_i y_i \log[\mathrm{E}(Y_i)] - \mathrm{E}(Y_i), \quad \frac{\partial \log \mathcal{L}[\mathrm{E}(\vec{Y})]}{\partial [\mathrm{E}(Y_i)]} = \frac{y_i}{\mathrm{E}(Y_i)} - 1$$

若 $\mathrm{E}(\vec{Y})$ 不受 $e^{(a+bx_i)}$ 约束，$\mathcal{L}[\mathrm{E}(\vec{Y})]$ 在 $\mathrm{E}(\vec{Y}) = \vec{y}$ 时取最大。

定义预测值 $\mathrm{E}(\vec{Y})$ 到观测值 $\vec{y}$ 的“距离” $\mathcal{D}$

- 令其越小代表两者符合得越好，取 0 时代表两者“相等”。

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left[\log \underbrace{\mathcal{L}[\mathrm{E}(\vec{Y}) = \vec{y}]}_{\text{似然函数的最大取值}} - \log \mathcal{L}[\mathrm{E}(\vec{Y})]\right]$$

$$= 2\left[\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right]$$

通常再乘上系数 2，动机来自正态分布（下页）。

对数似然距离

$$\log \mathcal{L}[\mathrm{E}(\vec{Y})] = \sum_i y_i \log[\mathrm{E}(Y_i)] - \mathrm{E}(Y_i), \quad \frac{\partial \log \mathcal{L}[\mathrm{E}(\vec{Y})]}{\partial [\mathrm{E}(Y_i)]} = \frac{y_i}{\mathrm{E}(Y_i)} - 1$$

若 $\mathrm{E}(\vec{Y})$ 不受 $e^{(a+bx_i)}$ 约束，$\mathcal{L}[\mathrm{E}(\vec{Y})]$ 在 $\mathrm{E}(\vec{Y}) = \vec{y}$ 时取最大。

定义预测值 $\mathrm{E}(\vec{Y})$ 到观测值 $\vec{y}$ 的“距离” $\mathcal{D}$

- 令其越小代表两者符合得越好，取 0 时代表两者“相等”。

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left[\log \underbrace{\mathcal{L}[\mathrm{E}(\vec{Y}) = \vec{y}]}_{\text{似然函数的最大取值}} - \log \mathcal{L}[\mathrm{E}(\vec{Y})]\right]$$

$$= 2\left[\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right]$$

通常再乘上系数 2 ，动机来自正态分布（下页）。

对数似然距离

$$\log \mathcal{L}[\mathrm{E}(\vec{Y})] = \sum_i y_i \log[\mathrm{E}(Y_i)] - \mathrm{E}(Y_i), \quad \frac{\partial \log \mathcal{L}[\mathrm{E}(\vec{Y})]}{\partial [\mathrm{E}(Y_i)]} = \frac{y_i}{\mathrm{E}(Y_i)} - 1$$

若 $\mathrm{E}(\vec{Y})$ 不受 $e^{(a+bx_i)}$ 约束，$\mathcal{L}[\mathrm{E}(\vec{Y})]$ 在 $\mathrm{E}(\vec{Y}) = \vec{y}$ 时取最大。

定义预测值 $\mathrm{E}(\vec{Y})$ 到观测值 $\vec{y}$ 的“距离” $\mathcal{D}$

- 令其越小代表两者符合得越好，取 0 时代表两者“相等”。

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left[\log \underbrace{\mathcal{L}[\mathrm{E}(\vec{Y}) = \vec{y}]}_{\text{似然函数的最大取值}} - \log \mathcal{L}[\mathrm{E}(\vec{Y})]\right]$$

$$= 2\left[\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right]$$

通常再乘上系数 2 ，动机来自正态分布（下页）。

残差平方和与对数似然距离

考虑正态分布的对数似然距离

$$\mathcal{D}[\mathrm{E}(\vec{Y}),\vec{y}]=2\{\underbrace{\log\mathcal{L}[\mathrm{E}(\vec{Y})=\vec{y}]}_{(\vec{y}-\vec{y})^2/(2\sigma^2)=0}-\log\mathcal{L}[\mathrm{E}(\vec{Y})]\}=\sum_i[\mathrm{E}(\vec{Y_i})-y_i]^2/\sigma^2$$

正是线性回归使用的残差平方和。

启示

泊松分布的“对数似然距离”，可看作是“残差平方和”的推广。

$$\hat{a},\hat{b}=\arg\max_{a,b}\log\mathcal{L}_i(a,b)=\arg\min_{a,b}\mathcal{D}(e^{a+bx_i},y_i)$$

残差平方和与对数似然距离

考虑正态分布的对数似然距离

$$\mathcal{D}[\mathrm{E}(\vec{Y}),\vec{y}]=2\{\underbrace{\log\mathcal{L}[\mathrm{E}(\vec{Y})=\vec{y}]}_{(\vec{y}-\vec{y})^2/(2\sigma^2)=0}-\log\mathcal{L}[\mathrm{E}(\vec{Y})]\}=\sum_i[\mathrm{E}(\vec{Y}_i)-y_i]^2/\sigma^2$$

正是线性回归使用的残差平方和。

启示

泊松分布的“对数似然距离”，可看作是“残差平方和”的推广。

$$\hat{a},\hat{b}=\arg\max_{a,b}\log\mathcal{L}_i(a,b)=\arg\min_{a,b}\mathcal{D}(e^{a+bx_i},y_i)$$

残差平方和与对数似然距离

考虑正态分布的对数似然距离

$$\mathcal{D}[\mathrm{E}(\vec{Y}),\vec{y}]=2\{\underbrace{\log\mathcal{L}[\mathrm{E}(\vec{Y})=\vec{y}]}_{(\vec{y}-\vec{y})^2/(2\sigma^2)=0}-\log\mathcal{L}[\mathrm{E}(\vec{Y})]\}=\sum_i[\mathrm{E}(\vec{Y_i})-y_i]^2/\sigma^2$$

正是线性回归使用的残差平方和。

启示

泊松分布的“对数似然距离”，可看作是“残差平方和”的推广。

$$\hat{a},\hat{b}=\arg\max_{a,b}\log\mathcal{L}_i(a,b)=\arg\min_{a,b}\mathcal{D}(e^{a+bx_i},y_i)$$

对数似然距离的直观理解

$$\sum_i [(a + bx_i) - y_i]^2$$

残差平方和：

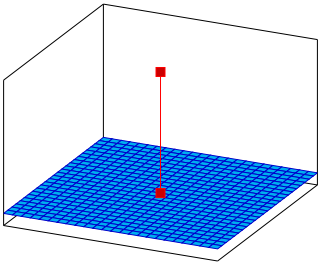


图: 线性回归

$$\mathcal{D}(e^{a+bx_i}, y_i)$$

对数似然距离，非线性：

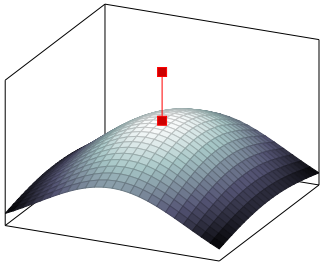


图: 泊松回归

使用最小二乘法，可解析求解线性回归模型，能否推广到泊松回归?

对数似然距离的直观理解

$$\sum_i [(a+bx_i)-y_i]^2$$

残差平方和：

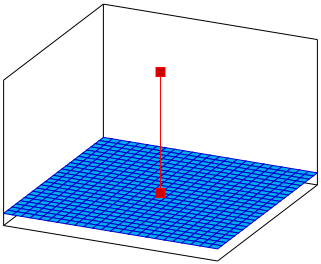


图: 线性回归

$$\mathcal{D}(e^{a+bx_i}, y_i)$$

对数似然距离，非线性：

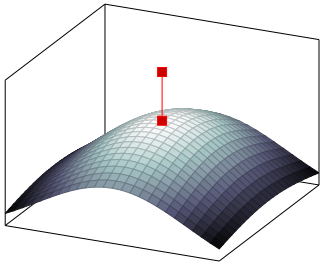


图: 泊松回归

使用最小二乘法，可解析求解线性回归模型，能否推广到泊松回归?

对数似然距离的直观理解

$$\sum_i [(a + bx_i) - y_i]^2$$

残差平方和：

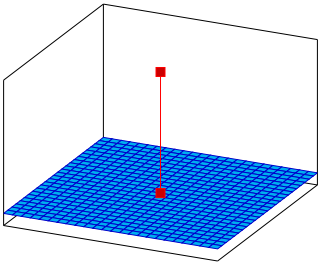


图: 线性回归

$$\mathcal{D}(e^{a+bx_i}, y_i)$$

对数似然距离，非线性：

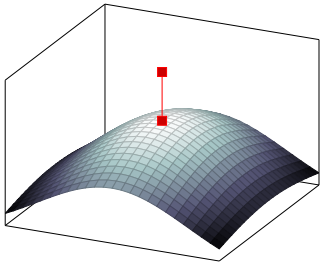


图: 泊松回归

使用最小二乘法，可解析求解线性回归模型，能否推广到泊松回归?

模型求解

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a}, \hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$：

$$\hat{a}, \hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a}, \hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$：

$$\hat{a}, \hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a}, \hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$:

$$\hat{a}, \hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a},\hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$:

$$\hat{a},\hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)}[y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b}\left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)}[y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)}[y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b}[y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a},\hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$:

$$\hat{a},\hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a},\hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$:

$$\hat{a},\hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

求解数似然函数距离的极小值

以最小对数似然距离求 $\hat{b}$：

$$\hat{a},\hat{b} = \arg\min_{a,b} \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]$$

$$\Rightarrow 0 = \frac{\partial \log \mathcal{D}[\mathrm{E}(Y_i|a,b), y_i]}{\partial b}$$

$$= \sum_i \left[-\frac{y_i}{\mathrm{E}(Y_i)} + 1 \right] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b}$$

构造最小二乘法求 $\hat{b}$:

$$\hat{a},\hat{b} = \arg\min_{a,b} \underbrace{\sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2}_{\text{加权的残差平方和}}$$

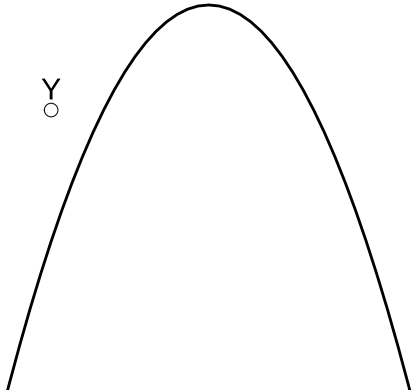
$$\Rightarrow 0 = \frac{\partial}{\partial b} \left\{ \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i|a,b)]^2 \right\}$$

$$\Longrightarrow 0 = \sum_i \frac{1}{\mathrm{E}(Y_i)} [y_i - \mathrm{E}(Y_i)] \frac{\partial \mathrm{E}(Y_i|a,b)}{\partial b} = -\frac{1}{2} \sum_i \frac{1}{\mathrm{E}(Y_i)} \frac{\partial}{\partial b} [y_i - \mathrm{E}(Y_i|a,b)]^2$$

- 两者解相同，对数似然距离与残差平方和等价。
- 当权重 $1/\mathrm{E}(Y_i)$ 被看作常数时，线性回归与泊松回归等价。

变权迭代最小二乘法

$$a_0, b_0 \to \mathrm{E}(Y_i) = a_0 + b_0 x_i \xrightarrow{\text{求解线性回归}} a_1, b_1 \to \mathrm{E}(Y_i) = a_1 + b_1 x_i \cdots$$

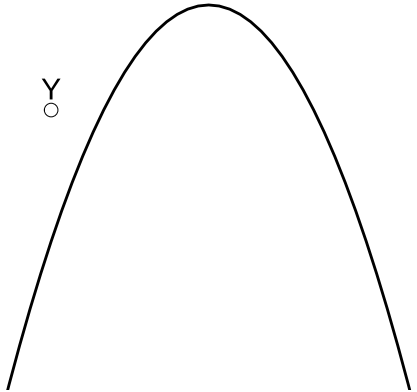


要寻找曲面上距离已知点 Y 最近的点，

1. 在曲线上猜一个位置 $B_0, i \leftarrow 0$；
2. 从 Y 向 B_i 处的切空间投影得 B'；
3. 令 $B_{i+1} \leftarrow B', i \leftarrow i+1$；
4. 若 $B_{i+1} = B_i$ 则得解，
   否则回到第 2 步。

变权迭代最小二乘法

$$a_0, b_0 \to \mathrm{E}(Y_i) = a_0 + b_0 x_i \xrightarrow{\text{求解线性回归}} a_1, b_1 \to \mathrm{E}(Y_i) = a_1 + b_1 x_i \cdots$$



要寻找曲面上距离已知点 Y 最近的点，

1. 在曲线上猜一个位置 $B_0, i \leftarrow 0$；
2. 从 Y 向 B_i 处的切空间投影得 B'；
3. 令 $B_{i+1} \leftarrow B', i \leftarrow i + 1$；
4. 若 $B_{i+1} = B_i$ 则得解，
   否则回到第 2 步。

泊松回归求解辐射量

- 信号数 N，依赖于样品辐射强度 I 和探头距离 r。

$$N = \frac{I}{r^s}$$

$$\log N = \log I - s \log r$$

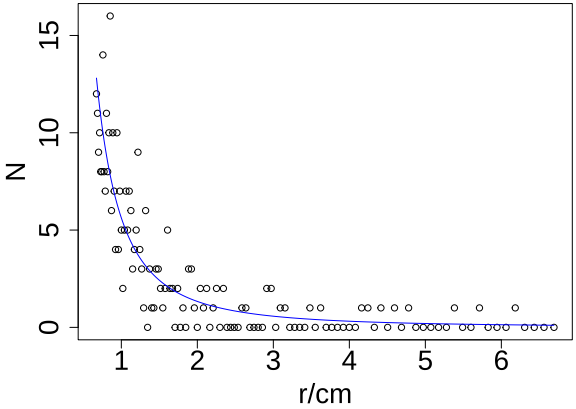
	$\log I$	s
线性回归	1.726 ± 0.067	-1.092 ± 0.067
泊松回归	1.727 ± 0.057	-2.08 ± 0.13
参考值	1.792	-2

泊松回归求解辐射量

- 信号数 N，依赖于样品辐射强度 I 和探头距离 r。

$$N = \frac{I}{r^s}$$

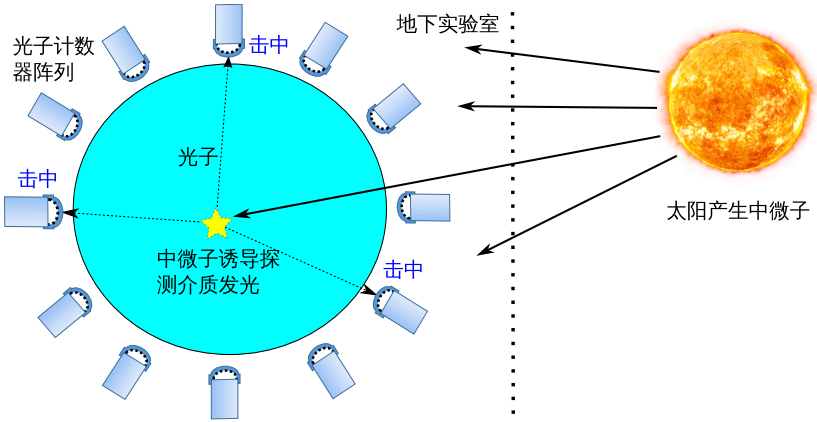
$$\log N = \log I - s \log r$$



	$\log I$	s
线性回归	1.726 ± 0.067	-1.092 ± 0.067
泊松回归	1.727 ± 0.057	-2.08 ± 0.13
参考值	1.792	-2

中微子案例

泊松回归在实验物理（中微子望远镜）的应用

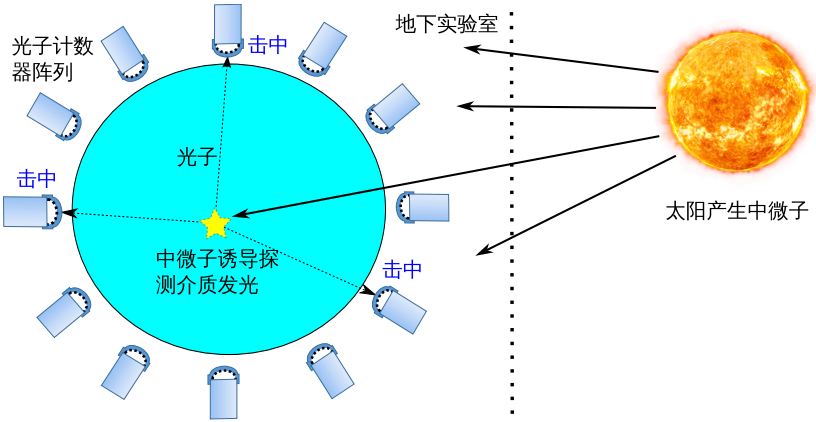


上万个计数器协同工作，测量中微子能量，定出中微子位置。

- 运用泊松回归方法实现测量精度的大幅提升。

《原子核仪器与物理学研究方法》期刊 NIMA Volume 1057, December 2023, 168692

泊松回归在实验物理（中微子望远镜）的应用



上万个计数器协同工作，测量中微子能量，定出中微子位置。

- 运用泊松回归方法实现测量精度的大幅提升。

《原子核仪器与物理学研究方法》期刊 NIMA Volume 1057, December 2023, 168692

刻度的泊松回归模型

$$N_i \sim \pi[\mathrm{E}(N_i)]$$

在中微子探测问题中，考虑中微子在介质中的反应位置 $\vec{r}$。需要获得 Y_i 与 $\vec{r}$ 的关系，当反应点与 $\vec{r}_i$ 处的 PMT 距离较远时，可用平方反比近似。

$$\mathrm{E}(N_i) = \frac{I}{(\vec{r}_i - \vec{r})^s}$$

如果再考虑介质对闪烁光的吸收，吸收长度为 L_0。

$$\mathrm{E}(N_i) = \frac{I}{|\vec{r}_i - \vec{r}|^s} e^{-\frac{|\vec{r}_i - \vec{r}|}{L_0}}$$

取 log 成为线性关系，

$$\log \mathrm{E}(N_i) = \log I - s \log(|\vec{r}_i - \vec{r}|) - \frac{1}{L_0}|\vec{r}_i - \vec{r}|$$

在探测器刻度中，$\vec{r}$ 已知。通过泊松回归获得 $\log I, s, \frac{1}{L_0}$。

刻度的泊松回归模型

$$N_i \sim \pi[\mathrm{E}(N_i)]$$

在中微子探测问题中，考虑中微子在介质中的反应位置 $\vec{r}$。需要获得 Y_i 与 $\vec{r}$ 的关系，当反应点与 $\vec{r}_i$ 处的 PMT 距离较远时，可用平方反比近似。

$$\mathrm{E}(N_i) = \frac{I}{(\vec{r}_i - \vec{r})^s}$$

如果再考虑介质对闪烁光的吸收，吸收长度为 L_0。

$$\mathrm{E}(N_i) = \frac{I}{|\vec{r}_i - \vec{r}|^s} e^{-\frac{|\vec{r}_i - \vec{r}|}{L_0}}$$

取 log 成为线性关系，

$$\log \mathrm{E}(N_i) = \log I - s \log(|\vec{r}_i - \vec{r}|) - \frac{1}{L_0}|\vec{r}_i - \vec{r}|$$

在探测器刻度中，$\vec{r}$ 已知。通过泊松回归获得 $\log I, s, \frac{1}{L_0}$。

刻度的泊松回归模型

$$N_i \sim \pi[\mathrm{E}(N_i)]$$

在中微子探测问题中，考虑中微子在介质中的反应位置 $\vec{r}$。需要获得 Y_i 与 $\vec{r}$ 的关系，当反应点与 $\vec{r}_i$ 处的 PMT 距离较远时，可用平方反比近似。

$$\mathrm{E}(N_i) = \frac{I}{(\vec{r}_i - \vec{r})^s}$$

如果再考虑介质对闪烁光的吸收，吸收长度为 L_0。

$$\mathrm{E}(N_i) = \frac{I}{|\vec{r}_i - \vec{r}|^s} e^{-\frac{|\vec{r}_i - \vec{r}|}{L_0}}$$

取 log 成为线性关系，

$$\log \mathrm{E}(N_i) = \log I - s \log(|\vec{r}_i - \vec{r}|) - \frac{1}{L_0}|\vec{r}_i - \vec{r}|$$

在探测器刻度中，$\vec{r}$ 已知。通过泊松回归获得 $\log I, s, \frac{1}{L_0}$。

事例重建的泊松回归模型

重建中，中微子事例的位置 $\vec{r}$ 与能量 E 是待求量。

$$\log \mathrm{E}(N_i) = \log I - s \log(|\vec{r}_i - \vec{r}|) - \frac{1}{L_0}|\vec{r}_i - \vec{r}|$$

1. 对于假定的位置 $\vec{r}$，基于刻度标定的 s, L_0，回归 I 得到事例能量；
2. $\hat{I}$ 替换 I 计算回归的似然函数 $\mathcal{L}(\vec{r})$；
3. 利用最大似然法求得 $\hat{\vec{r}} = \arg\max_{\vec{r}} \mathcal{L}(\vec{\nabla})$ 。

思考（开放问题，取得解者 +5%）

有可能把 $\hat{\vec{r}}$ 的求解过程化为线性回归问题吗？

Ghost Hunter 2024

Ghost Hunter 2024 一般性地求解 $\mathrm{E}[N_i(\vec{r})]$。

- 再加上时间维度的信息，$\mathrm{E}[N_i(\vec{r}, t)]$。
- 可考虑使用回归方法取得函数形式，例如对时间取小区间。
- 难点：加入了时间之后，如何线性化表示 $N_i(\vec{r}, t)$ 函数关系？
  - Ghost Hunter 上的百花齐放，各显神通时刻。
  - 欢迎加入战斗！

总结推广

总结

1. 把线性回归中的正态假设替换为泊松假设，得到泊松回归。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i.$$

将简洁有效的线性回归方法范围推广到计数问题。

2. 泊松对数似然距离，是正态残差平方和的推广：

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left\{\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right\}.$$

3. 变权迭代最小二乘法把线性回归的框架应用于泊松回归：

$$a_0, b_0 \to \mathrm{E}(Y_i) = a_0 + b_0 x_i \xrightarrow{\text{求解线性回归}} a_1, b_1 \to \mathrm{E}(Y_i) = a_1 + b_1 x_i \cdots$$

启示：用线性方法可以高效解决看似非线性的问题。

总结

1. 把线性回归中的正态假设替换为泊松假设，得到泊松回归。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i.$$

将简洁有效的线性回归方法范围推广到计数问题。

2. 泊松对数似然距离，是正态残差平方和的推广：

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left\{\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right\}.$$

3. 变权迭代最小二乘法把线性回归的框架应用于泊松回归：

$$a_0, b_0 \to \mathrm{E}(Y_i) = a_0 + b_0 x_i \xrightarrow{\text{求解线性回归}} a_1, b_1 \to \mathrm{E}(Y_i) = a_1 + b_1 x_i \cdots$$

启示：用线性方法可以高效解决看似非线性的问题。

总结

1. 把线性回归中的正态假设替换为泊松假设，得到泊松回归。

$$Y_i \sim \pi[\mathrm{E}(Y_i)], \log \mathrm{E}(Y_i) = a + bx_i.$$

将简洁有效的线性回归方法范围推广到计数问题。

2. 泊松对数似然距离，是正态残差平方和的推广：

$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left\{\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right\}.$$

3. 变权迭代最小二乘法把线性回归的框架应用于泊松回归：

$$a_0, b_0 \to \mathrm{E}(Y_i) = a_0 + b_0 x_i \xrightarrow{\text{求解线性回归}} a_1, b_1 \to \mathrm{E}(Y_i) = a_1 + b_1 x_i \cdots$$

启示：用线性方法可以高效解决看似非线性的问题。

广义线性回归

泊松回归是广义线性回归的一个特例。

- 常见的广义线性回归有
  - 伽马回归：观测量是正实数
  - 逻辑回归：观测量是 $[0, 1]$ 区间内的实数
  - 二项回归：观测是是 $0, 1 \cdots, N$ 的整数
- 它们都可以通过变权迭代最小二乘法有效解出。