广义线性回归

续本达

清华大学 工程物理系

2024-12-04

复习

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：

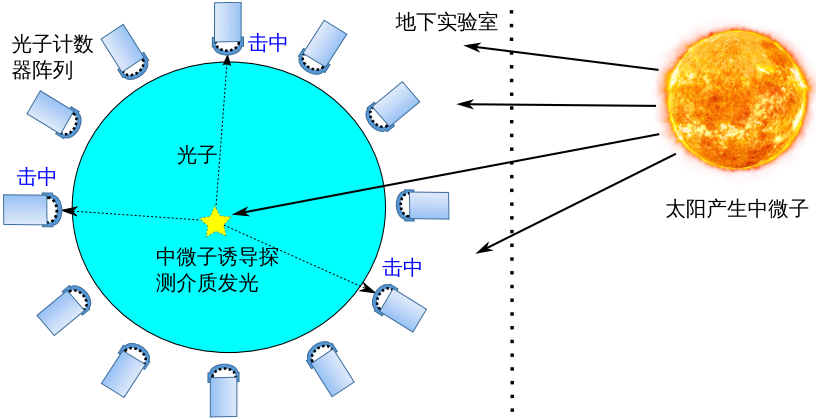
$$\mathcal{D}[\mathrm{E}(\vec{Y}), \vec{y}] = 2\left\{\sum_i y_i \log \frac{y_i}{\mathrm{E}(Y_i)} - [y_i - \mathrm{E}(Y_i)]\right\}.$$

1. 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。

泊松回归在实验物理（中微子望远镜）的应用

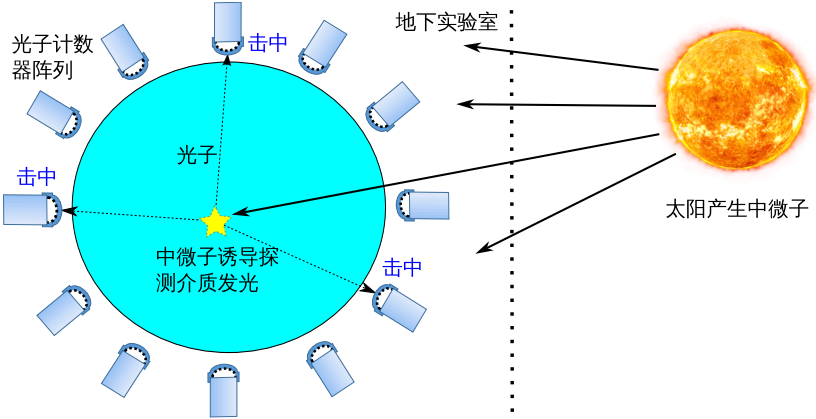


上万个计数器协同工作，测量中微子能量，定出中微子位置。

- 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

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！

指数离散分布族

定义回顾

指数分布族

一族的分布，它们的概率密度函数或分布律记为 $f(x|\vec{\theta})$ ，如果有

$$f(x|\vec{\theta}) = \exp\left[\sum_{i=1}^{s} \eta_i(\vec{\theta})T_i(x) - B(\vec{\theta})\right] h(x)$$

的形式，则称为 s 维指数分布族 。

如果把 $\eta_i(\vec{\theta})$ 当作自变量反解 θ，那么形式有进一步的化简。

指数分布族标准形式

$$f(x|\vec{\eta}) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\vec{\eta})\right] h(x)$$

其中 $\eta_i = \eta_i(\vec{\theta})$, $A[\vec{\eta}(\vec{\theta})] = B(\vec{\theta})$。

指数离散分布族 Exponential Dispersion Family

在指数分布族上再配一项 $\phi > 0$。

$$f(x|\vec{\eta}) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\vec{\eta})\right] h(x)$$

变为

$$f(x|\phi, \vec{\eta}) = \exp\left\{\sum_{i=1}^{s} \frac{1}{\phi}\left[\eta_i T_i(x) - A(\vec{\eta})\right]\right\} h(\phi, x)$$

当 ϕ 取 1 时，Exponential Dispersion Family 就是 Exponential Family。

期望与方差

- 期望：$\mathrm{E}(X) = A'(\eta)$ 不变，记为 μ
- 方差：$\mathrm{Var}(X) = A''(\eta) \to \mathrm{Var}(X) = \phi A''(\eta)$
  - 所以 ϕ 称为 离散度参数 dispersion parameter。

方差与期望的相关性

- 期望：$\mathrm{E}(X) = \mu = A'(\eta)$
- 方差：$\mathrm{Var}(X) = \phi A''(\eta)$ 说明 $A(\eta)'' > 0$
- 说明 $\mu(\eta) = A'(\eta)$ 是单调函数，η 与 μ 之间有双射。

$$A'^{-1}(\mu) = \eta$$

方差函数

把 $A''(\eta)$ 换为 $A''(A'^{-1}(\mu))$ 定义为 $V(\mu)$，称 方差函数 。那么

$$\mathrm{Var}(X) = \phi V(\mu)$$

常见的指数离散分布族成员

名称	$V(\mu)$	η	ϕ	$d(x,\mu)$
正态	1	μ	σ^2	
二项	$\mu(1-\mu)$	$\log\frac{\mu}{1-\mu}$	$\frac{1}{N}$	
泊松	μ	$\log\mu$	1	
伽马	μ^2	$-\frac{1}{\mu}$	ϕ	
负二项	$\mu+\frac{\mu^2}{k}$	$\log\frac{\mu}{\mu+k}$	1	
逆高斯	μ^2	$-\frac{1}{2\mu^2}$	ϕ	
Tweedie	μ^ξ	$\frac{\mu^{1-\xi}}{1-\xi}$	ϕ	

对数似然距离记为 $d(x,\mu)$ ，近似地

$$\frac{d(x,\mu)}{\phi}\sim\chi^2(1)$$

- F-检验 与 t-检验都适用。

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
  - 负二项回归：观测量是非负整数，比泊松回归多一个参数
  - 伽马回归：观测量是正实数
    - 或者使用逆高斯分布。
  - 逻辑回归：观测量是 $[0, 1]$ 区间内的实数
  - 二项回归：观测是 $0, 1 \cdots, N$ 的整数
    - 关注成功次数占总数的比例。
  - Tweedie 回归：
- 它们都可以通过变权迭代最小二乘法有效解出。

案例二刷

方差分析

- 二项回归

北京降雨

- Tweedie 回归

泊松回归例子

- 尝试对 r10500.h5 进行泊松回归
- 辅助文件 geo.csv，pmtdata.csv

```
library(rhdf5)
library(dplyr)
PEs <- h5read("lm-examples/r10500.h5", "PETruth")
PEs %>% group_by(EventID, ChannelID) %>% summarise(nPE=n()) -> nPE
PMT <- read.table("lm-examples/pmtdata.csv", col.names=c("ChannelID", "type", "QE"))
nPE %>% inner_join(PMT) -> d
lm0.d <- glm(nPE ~ type - 1, d, family = poisson())
summary(lm0.d)
```

数据集下载

- lm-examples.tar.gz 线性回归样例集

```
wget http://hep.tsinghua.edu.cn/~orv/pd/lm-examples.tar.gz
tar -xf lm-examples.tar.gz
```

泊松回归结果解读

回归结果输出为

```
Call:
glm(formula = nPE ~ type - 1, family = poisson(), data = d)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
typeHamamatsu  3.3682078  0.0005870  5738.4   <2e-16 ***
typeHighQENNVT 3.3916647  0.0003988  8504.3   <2e-16 ***
typeHZC        0.3696733  0.0018707   197.6   <2e-16 ***
typeNNVT       3.2986176  0.0009531  3461.0   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 55350655  on 549674  degrees of freedom
Residual deviance:  5729167  on 549670  degrees of freedom
AIC: 7915838

Number of Fisher Scoring iterations: 5
```