自助法

续本达

清华大学 工程物理系

2023-12-11

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。

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。

非参数估计

参数估计到非参数估计

- 非参数不如叫“无穷推断”，模型有无穷维。
  - “模型”是“解集”的别称。
- 无穷维的好处：简化假设。
  - 坏处：解集太大。

例子

- 估计分布函数
- 概率密度函数：使用直方图估计概率密度函数
- 估计相关关系

参考书 Larry Wasserman, All of nonparametric statistics

参数估计到非参数估计

- 非参数不如叫“无穷推断”，模型有无穷维。
  - “模型”是“解集”的别称。
- 无穷维的好处：简化假设。
  - 坏处：解集太大。

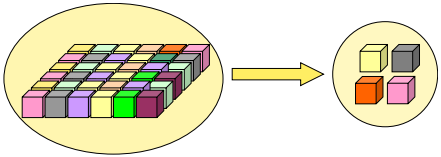
例子

- 估计分布函数
- 概率密度函数：使用直方图估计概率密度函数
- 估计相关关系

参考书 Larry Wasserman, All of nonparametric statistics

回顾总体与样本

样本从总体中抽取，并作为总体代表的一部分总体单位的集合体。



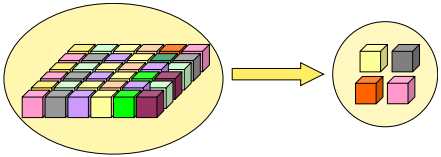
样本取自总体，不唯一

抽样分布的 3 个要素

1. 总体
2. 抽样方法
   本课程的范围中，都是简单随机样本
3. 统计量
   是样本的函数，是随机变量，其分是抽样分布

回顾总体与样本

样本从总体中抽取，并作为总体代表的一部分总体单位的集合体。



样本取自总体，不唯一

抽样分布的 3 个要素

1. 总体
2. 抽样方法
   本课程的范围中，都是简单随机样本
3. 统计量
   是样本的函数，是随机变量，其分是抽样分布

使用过的抽样分布

- 正态分布、t 分布、χ^2 分布、F 分布
  - 参数估计、区间估计、假设检验、回归分析、方差分析
- 如果抽样分布未知怎么办？例如中位数
  - 区间估计、假设检验如何做？
  - 提示：数据蕴含了抽样分布的信息。

使用过的抽样分布

- 正态分布、t 分布、χ^2 分布、F 分布
  - 参数估计、区间估计、假设检验、回归分析、方差分析
- 如果抽样分布未知怎么办？例如中位数
  - 区间估计、假设检验如何做？
  - 提示：数据蕴含了抽样分布的信息。

使用过的抽样分布

- 正态分布、t 分布、χ^2 分布、F 分布
  - 参数估计、区间估计、假设检验、回归分析、方差分析
- 如果抽样分布未知怎么办？例如中位数
  - 区间估计、假设检验如何做？
  - 提示：数据蕴含了抽样分布的信息。

自助法

语源

Why can not a man lift himself by pulling up on his bootstraps?

《英雄无泪》古龙

小高没有被她拖下去，反而又向上拔起，以右脚垫左脚，借力使力，又向上拔起丈余，就看见窄巷两边的短墙后，都有一个人分别向左右两方窜出，身手都极矫健，轻功都不弱。

《白眉大侠》单田芳

一下上不去，他身子蹿起来，在一丈五六尺的时候，左脚一踩右脚的脚面，这就换了一下气。‘噌！’然后右脚又踩了一下左脚尖，又拔起了七八尺高，这才到了台上”

语源

Why can not a man lift himself by pulling up on his bootstraps?

《英雄无泪》古龙

小高没有被她拖下去，反而又向上拔起，以右脚垫左脚，借力使力，又向上拔起丈余，就看见窄巷两边的短墙后，都有一个人分别向左右两方窜出，身手都极矫健，轻功都不弱。

《白眉大侠》单田芳

一下上不去，他身子蹿起来，在一丈五六尺的时候，左脚一踩右脚的脚面，这就换了一下气。‘噌！’然后右脚又踩了一下左脚尖，又拔起了七八尺高，这才到了台上”

语源

Why can not a man lift himself by pulling up on his bootstraps?

《英雄无泪》古龙

小高没有被她拖下去，反而又向上拔起，以右脚垫左脚，借力使力，又向上拔起丈余，就看见窄巷两边的短墙后，都有一个人分别向左右两方窜出，身手都极矫健，轻功都不弱。

《白眉大侠》单田芳

一下上不去，他身子蹿起来，在一丈五六尺的时候，左脚一踩右脚的脚面，这就换了一下气。‘噌!’然后右脚又踩了一下左脚尖，又拔起了七八尺高，这才到了台上”

估计量

估计量是用于近似总体的未知参数的统计量

代换原则

当总体参数未知时，使用它的估计量替代它

- $\mu \to \overline{X}$
- $\sigma^2 \to S^2$

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

能否代换一个总体的“估计量”？自助 （bootstrap）

1. 总体
2. 抽样
3. 统计量

估计量

估计量是用于近似总体的未知参数的统计量

代换原则

当总体参数未知时，使用它的估计量替代它

- $\mu \to \overline{X}$
- $\sigma^2 \to S^2$

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

能否代换一个总体的“估计量”？自助 （bootstrap）

1. 总体
2. 抽样
3. 统计量

估计量

估计量是用于近似总体的未知参数的统计量

代换原则

当总体参数未知时，使用它的估计量替代它

- $\mu \to \overline{X}$
- $\sigma^2 \to S^2$

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

能否代换一个总体的“估计量”？自助 （bootstrap）

1. 总体
2. 抽样
3. 统计量

替换总体：自助法

1. 把样本 F_n 看作总体 F 的估计
2. 抽样
3. 统计量

样本看作总体的估计

用经验分布函数 F_n 估计总体分布函数 F

经验分布函数

$x_1, x_2, \cdots, x_n$ 是来自分布函数为 $F(x)$ 总体 X 的样本观察值。 X 的经验分布函数 $F_n(x)$ 定义为

$$F_n(x) = \frac{\#(x_i \le x)}{n}$$

从 F_n 抽样可以看作是对样本 $x_1, x_2, \cdots, x_n$ 进行可放回的重新抽样，称作 bootstrap 样本，记为 $x_1^*, x_2^*, \cdots, x_n^*$ 或 $\vec{x}^*$。

替换总体：自助法

1. 把样本 F_n 看作总体 F 的估计
2. 抽样
3. 统计量

样本看作总体的估计

用经验分布函数 F_n 估计总体分布函数 F

经验分布函数

$x_1, x_2, \cdots, x_n$ 是来自分布函数为 $F(x)$ 总体 X 的样本观察值。 X 的经验分布函数 $F_n(x)$ 定义为

$$F_n(x) = \frac{\#(x_i \le x)}{n}$$

从 F_n 抽样可以看作是对样本 $x_1, x_2, \cdots, x_n$ 进行可放回的重新抽样，称作 bootstrap 样本，记为 $x_1^*, x_2^*, \cdots, x_n^*$ 或 $\vec{x}^*$。

替换总体：自助法

1. 把样本 F_n 看作总体 F 的估计
2. 抽样
3. 统计量

样本看作总体的估计

用经验分布函数 F_n 估计总体分布函数 F

经验分布函数

$x_1, x_2, \cdots, x_n$ 是来自分布函数为 $F(x)$ 总体 X 的样本观察值。 X 的经验分布函数 $F_n(x)$ 定义为

$$F_n(x) = \frac{\#(x_i \le x)}{n}$$

从 F_n 抽样可以看作是对样本 $x_1, x_2, \cdots, x_n$ 进行可放回的重新抽样，称作 bootstrap 样本，记为 $x_1^*, x_2^*, \cdots, x_n^*$ 或 $\vec{x}^*$。

自助法

使用一系列 $\vec{x}_b^*$ ，得到相应的 $\hat{\theta}_b^* = \hat{\theta}(\vec{x}_b^*)$。定义

$$\overline{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

为 θ 的 bootstrap 估计

定理 (bootstrap 中心极限定理)

$$\overline{\theta^*} - \hat{\theta} \xrightarrow{L} \mathrm{E}(\hat{\theta}) - \theta$$

代换原则，当总体 F 换为经验分布 F_n 时，$\mathrm{E}(\hat{\theta}) \to \overline{\theta^*}$ ， $\theta \to \hat{\theta}$

收敛速度

$\overline{\theta^*} - \hat{\theta}$ 收敛到 $\mathrm{E}(\hat{\theta}) - \theta$ 的速度比正态分布中心极限定理预测得更快，称为自助法的 二次修正

自助法

使用一系列 $\vec{x}_b^*$ ，得到相应的 $\hat{\theta}_b^* = \hat{\theta}(\vec{x}_b^*)$。定义

$$\overline{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

为 θ 的 bootstrap 估计

定理 (bootstrap 中心极限定理)

$$\overline{\theta^*} - \hat{\theta} \xrightarrow{P} \mathrm{E}(\hat{\theta}) - \theta$$

代换原则，当总体 F 换为经验分布 F_n 时，$\mathrm{E}(\hat{\theta}) \to \overline{\theta^*}$ ， $\theta \to \hat{\theta}$

收敛速度

$\overline{\theta^*} - \hat{\theta}$ 收敛到 $\mathrm{E}(\hat{\theta}) - \theta$ 的速度比正态分布中心极限定理预测得更快，称为自助法的 二次修正

自助法

使用一系列 $\vec{x}_b^*$ ，得到相应的 $\hat{\theta}_b^* = \hat{\theta}(\vec{x}_b^*)$。定义

$$\overline{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$$

为 θ 的 bootstrap 估计

定理 (bootstrap 中心极限定理)

$$\overline{\theta^*} - \hat{\theta} \xrightarrow{P} \mathrm{E}(\hat{\theta}) - \theta$$

代换原则，当总体 F 换为经验分布 F_n 时，$\mathrm{E}(\hat{\theta}) \to \overline{\theta^*}$ ， $\theta \to \hat{\theta}$

收敛速度

$\overline{\theta^*} - \hat{\theta}$ 收敛到 $\mathrm{E}(\hat{\theta}) - \theta$ 的速度比正态分布中心极限定理预测得更快，称为自助法的 二次修正

$\mathrm{E}(\hat{\theta}) \to \overline{\theta^*}$ 与蒙特卡罗方法

$$\mathrm{E}_F(\hat{\theta}) \to \mathrm{E}_{F_n}(\hat{\theta}) \xrightarrow{\text{Monte Carlo}} \overline{\theta^*}$$

- 计算 $\mathrm{E}_{F_n}(\hat{\theta})$ ，需要考虑所有的 n^n 种重抽样可能性。
- 计算量过大，使用了蒙特卡罗方法近似。
  从 $x_1, x_2, \cdots, x_n$ 进行 随机 重抽样

例：量子基金

自助法原则

将总体分布代换为经验分布后， 一切 都按照基本原理计算。

中位数

例：中位数的区间估计

量子基金过去 13 年的年回报率为

```
fund <- c(9.5, 21.1, 12.0, 10.2, 12.0, 21.1, 10.2, 18.2, 12.0, 9.5, 18.0, 10.2, 18.2)
```

求回报率中位数的区间估计

```
library(plyr)
m.bootstrap <- laply(1:10000, function(x) {
    s <- sample(fund, length(fund), replace = TRUE)
    median(s)
})
mean(m.bootstrap)
var(m.bootstrap)
sd(m.bootstrap)
```

```
[1] 12.81137
[1] 7.144108
[1] 2.672846
```

中位数

例：中位数的区间估计

量子基金过去 13 年的年回报率为

```
fund <- c(9.5, 21.1, 12.0, 10.2, 12.0, 21.1, 10.2, 18.2, 12.0, 9.5, 18.0, 10.2, 18.2)
```

求回报率中位数的区间估计

```
library(plyr)
m.bootstrap <- laply(1:10000, function(x) {
    s <- sample(fund, length(fund), replace = TRUE)
    median(s)
})
mean(m.bootstrap)
var(m.bootstrap)
sd(m.bootstrap)
```

```
[1] 12.81137
[1] 7.144108
[1] 2.672846
```

估计量偏差的估计

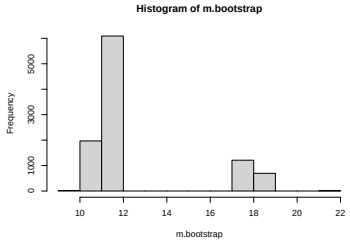
$$\overline{m^*} - \hat{m} \to \underbrace{\mathrm{E}(\hat{m}) - m}_{\text{偏差}}$$

bootstrap 估计量与经典估计量之差，估计了经典估计量的偏差。
经过 bootstrap 修正的估计量：

$$m \approx \mathrm{E}\,\hat{m} - (\overline{m^*} - \hat{m}) \approx 2\hat{m} - \overline{m^*}$$

区间估计


```
hist(m.bootstrap)
```



- 使用 $\hat{\theta}_b^*$ 的 bootstrap 样本分位数，进行区间估计

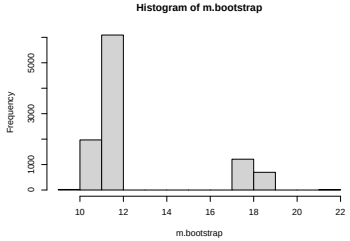
```
quantile(m.bootstrap, c(0.025, 0.975))
```

```
2.5% 97.5%
10.2  18.2
```

- 量子基金年收益中位数的 95% 置信水平的置信区间为 $(10.2, 18.2)$

区间估计


```
hist(m.bootstrap)
```



- 使用 $\hat{\theta}_b^*$ 的 bootstrap 样本分位数，进行区间估计

```
quantile(m.bootstrap, c(0.025, 0.975))
```

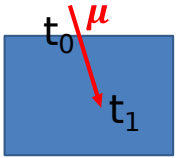
```
2.5% 97.5%
10.2  18.2
```

- 量子基金年收益中位数的 95% 置信水平的置信区间为 $(10.2, 18.2)$

例：缪子寿命

缪子寿命测量实验

缪子寿命



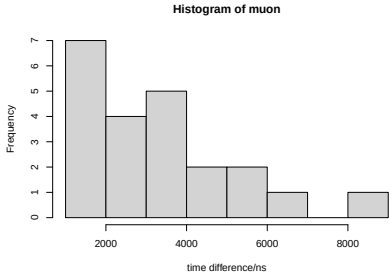
工物系的樊同学测量了一组停在探测器中的宇宙线缪子实验的数据，具体为缪子衰变时刻 t_1 与缪子进入探测器时刻 t_0 之差

```
muon <- c(5793, 1055, 6099, 3430, 3733, 4212, 1143, 1467, 5269, 2332, 1388, 1019, 2123, 3153, 3042,
          1425, 8977, 2686, 3307, 1260, 2577, 4419)
```

求缪子寿命的区间估计。

缪子寿命


```
hist(muon, xlab="time difference/ns")
```



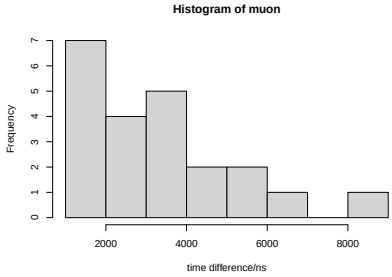
- 猜测是指数分布，因为取数的系统时间延迟而有平移，分布函数形式为

$$F(t|\tau, s) = 1 - e^{-(t-s)/\tau}$$

平移可能来自于信号线的延迟差异。

缪子寿命


```
hist(muon, xlab="time difference/ns")
```



- 猜测是指数分布，因为取数的系统时间延迟而有平移，分布函数形式为

$$F(t|\tau, s) = 1 - e^{-(t-s)/\tau}$$

平移可能来自于信号线的延迟差异。

构造统计量

以 t_i 代表实验测量值

$$\begin{cases} \hat{s} = \min t_i - \frac{1}{n}\hat{\tau} \\ \hat{\tau} = \bar{t} - \hat{s} \end{cases} \implies \hat{\tau} = \frac{\bar{t} - \min t_i}{1 - 1/n}$$

```
tau <- (mean(muon) - min(muon)) / (1-1/length(muon))
tau
```

```
[1] 2261.476
```

- 与 2.2 μs 的缪子寿命相符吗?

构造统计量

以 t_i 代表实验测量值

$$\begin{cases} \hat{s} = \min t_i - \frac{1}{n}\hat{\tau} \\ \hat{\tau} = \bar{t} - \hat{s} \end{cases} \implies \hat{\tau} = \frac{\bar{t} - \min t_i}{1 - 1/n}$$

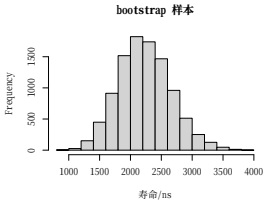
```
tau <- (mean(muon) - min(muon)) / (1-1/length(muon))
tau
```

```
[1] 2261.476
```

- 与 2.2 μs 的缪子寿命相符吗?

寿命的区间估计


```
library(plyr)
t.bootstrap <- laply(1:10000, function(x) {
    s <- sample(muon, length(muon), replace = TRUE)
    (mean(s) - min(s)) / (1-1/length(muon))
})
hist(t.bootstrap, main="bootstrap 样本", xlab="寿命/ns")
```




```
quantile(t.bootstrap, prob=c(0.05, 0.95))
```

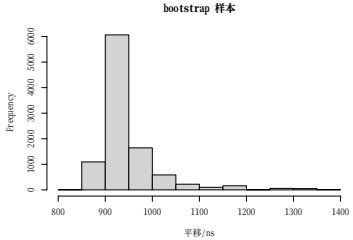
```
      5%      95%
1551.952 2949.176
```

- 测得 μ 子寿命 90% 置信区间为 $(1552, 2949)$ns

偏移量的 bootstrap

$$\hat{s} = \hat{\tau} - \bar{t}$$

```
s.bootstrap <- laply(1:10000, function(x) {
    s <- sample(muon, length(muon), replace = TRUE)
    mean(s) - (mean(s) - min(s)) / (1-1/length(muon))
})
hist(s.bootstrap, main="bootstrap 样本", xlab="平移/ns", bins=20)
```



偏移量的假设检验

$$H_0 : s = 0, H_1 : s > 0$$

$\hat{s}_b^* - \overline{s^*}$ 是 $\hat{s}_b^* - \mathrm{E}(\hat{s}) = \hat{s}_b^* - s$ 的估计，后者为原假设下的 bootstrap 采样。p 值的估计为

$$P(\hat{s}' \geq \hat{s}|H_0) \approx \frac{\#(\hat{s}_b^* - \overline{s^*} \geq \hat{s})}{B}$$

```
s.test <- s.bootstrap - mean(s.bootstrap) >= s
sum(s.test) / length(s.test)
```

```
[1] 0
```

- p 值的 bootstrap 估计为 0，拒绝原假设，应该保留 s 一项。

偏移量的假设检验

$$H_0 : s = 0, H_1 : s > 0$$

$\hat{s}_b^* - \overline{s^*}$ 是 $\hat{s}_b^* - \mathrm{E}(\hat{s}) = \hat{s}_b^* - s$ 的估计，后者为原假设下的 bootstrap 采样。p 值的估计为

$$P(\hat{s}' \geq \hat{s}|H_0) \approx \frac{\#(\hat{s}_b^* - \overline{s^*} \geq \hat{s})}{B}$$

```
s.test <- s.bootstrap - mean(s.bootstrap) >= s
sum(s.test) / length(s.test)
```

```
[1] 0
```

- p 值的 bootstrap 估计为 0，拒绝原假设，应该保留 s 一项。

回归分析

回归系数的置信区间

$$Y_i = a + bx_i + \epsilon_i$$

使用 bootstrap 估计

1. 对 (x, y) 数据集应用 bootstrap 方法重新抽样；
2. 对每一个抽样做回归分析，获得 a^*, b^* 等参数；
3. 使用 a^*, b^* 的分位数做区间估计。

对比：使用 t-分布 对 a, b 做区间估计。

总结

总结

bootstrap 方法的特点

1. 概念简洁
   不需要抽样分布，只需重复采样，方法不变
2. 计算量大
   计算机代替我们完成，要求我们熟练使用编程语言。推荐 R 语言。
3. 与传统方法配合使用
   bootstrap 使用计算机代替了对问题的直觉理解，戒“无脑滥用”。