权重的使用与加权最小二乘

慧航

2024年4月

加权平均

为了使得样本对总体的代表性更好，我们可能需要在做统计分析时对数据进行加权。

- 比如，在很多的抽样调查中，为了得到比较好的统计性质，通常抽样并非等概率的进行，而是针对某一群体，比如低收入家庭等，进行有重点地抽样。
- 此时，如果我们希望获得收入的总体均值，简单的计算样本平均会导致平均收入的低估。

Horvitz-Thompson估计量

为了解决这一问题，通常我们会使用Horvitz-Thompson估计量，也就是使用每个样本被抽中的概率π_i的倒数$1/\pi_i$作为权重，计算加权平均：

$$\bar{x}^w = \frac{1}{M}\sum_{i=1}^{N}\frac{1}{\pi_i}x_i$$

其中$M = \sum_{i=1}^{N}\frac{1}{\pi_i}$为总体中个体的数量。Horvitz-Thompson估计量为总体均值的无偏估计量。

逆概率加权

重新整理该估计量，有：

$$\bar{x}^w = \frac{1}{M}\sum_{i=1}^{N}\frac{1}{\pi_i}x_i = \frac{\sum_{i=1}^{N}\frac{1}{\pi_i}x_i}{\sum_{i=1}^{N}\frac{1}{\pi_i}} = \sum_{i=1}^{N}\left(\frac{\frac{1}{\pi_i}}{\sum_{i=1}^{N}\frac{1}{\pi_i}}\right)x_i \triangleq \sum_{i=1}^{N} w_i x_i$$

以上估计量也称为加权平均（weighted average），其中权重为概率的倒数，因而通常也被称为**逆概率加权**（**inverse probability weighting**）。

逆概率加权

概率的倒数可以简单理解为一个样本代表了总体中多少个个体，

- 如果一个样本被抽中的概率为万分之一，那么一个样本大概代表了一万个个体。
- 在实际的调查数据中，通常会使用“一个个体代表了多少个个体”这种形式来给出权重。

逆概率加权

Stata中的加权

在中国家庭金融调查（China Household Finance Survey）的数据（chfs_ind.dta）中，如果不使用权重，swgt变量指明了数据中1个人代表了总体中的多少人。在Stata中，有四种加权方式可供使用：

- fweight：频数权重，即如果我们观察到m个一模一样的观测，那么我们可以将这m个一模一样的观测合并为一个观测，并设其权重为m。
- aweight：分析权重，适用于加总的数据，比如我们使用的数据为每个省份的平均值，那么可以用省份的人口作为权重。权重在使用时会默认规范化所有的权重之和为N：$\sum_{i=1}^{N} w_i = N$。
- pweight：抽样权重，适用于抽样数据，权重为每个个体被抽中的概率的倒数，也就是我们刚刚提到的权重。
- iweight：重要性权重，Stata内部处理方法与aweight类似，区别在于使用iweight不会做规范化，适用于出于其他目的的加权。

逆概率加权

Stata中的加权

我们可以使用如下命令计算加权平均:

```
1 su labor_inc [aw=swgt]
```

加总数据的加权

加权平均的另一个应用是在加总的数据中。比如如果我们有每个城市的平均收入，为了计算全国的平均收入，我们可以按照如下计算：

$$\bar{x} = \frac{\sum_{c=1}^{C}(\bar{x}_c \times p_c)}{\sum_{c=1}^{C} p_c} = \sum_{c=1}^{C}\left(\frac{p_c}{\sum_{c=1}^{C} p_c} \times \bar{x}_c\right) \triangleq \sum_{c=1}^{C}(w_c \times \bar{x}_c)$$

实际上，以上计算方法也是一种逆概率加权：

- 由于对于城市数据而言，每个城市只有1条数据，从而$1/p_c$代表了每个城市c中一个个体被抽中的概率
- 从而根据逆概率加权的思想，权重应该为$1/(1/p_c)=p_c$，即使用人口数量进行加权。

加总数据的加权

使用城市平均计算全国平均

如果我们需要计算2010年全国人均公共图书册数，使用citydata.dta中的城市数据，我们分别计算了使用人口加权和不加权两种不同的均值：

```
use datasets/citydata.dta, clear
keep if year==2010
su v210
su v210 [aw=v4]
```

根据计算结果，未使用加权平均计算的人均公共图书册数约为4.89册，而使用人口加权后，得到的结果为5.27册，如果人口多的城市人均公共册数也多，那么后者对于全国人均公共图书册数的计算更加精确，而未加权的结果会低估全国的平均册数。

加权最小二乘

在线性回归中同样面临着需要加权的问题，此时可以使用加权最小二乘法，即最小化经过权重调整的误差平方和：

$$\min_{\beta} \sum_{i=1}^{N} \left[w_i \left(y_i - x_i' \beta \right)^2 \right]$$

从而得到：

$$\hat{\beta}^w = \left(\sum_{i=1}^{N} w_i x_i x_i' \right)^{-1} \left(\sum_{i=1}^{N} w_i x_i y_i \right)$$

其中w_i为权重。

加权最小二乘

在一些抽样调查数据中，可以使用逆概率加权的方法对估计量进行调整，即令$w_i = 1/\pi_i$，这与我们上面计算加权平均的方法是一样的。如果使用逆概率加权，我们可以将以上计算公式写为：

$$\hat{\beta}^w = \left(\frac{\sum_{i=1}^{N} \frac{1}{\pi_i} x_i x_i'}{\sum_{i=1}^{N} \frac{1}{\pi_i}} \right)^{-1} \left(\frac{\sum_{i=1}^{N} \frac{1}{\pi_i} x_i y_i}{\sum_{i=1}^{N} \frac{1}{\pi_i}} \right)$$

即将OLS公式中的两部分分别替换为了其无偏估计。

加权最小二乘

CHFS中的权重

在2002年中国家庭金融调查数据中，如果要比较不同教育程度的对数收入，可以使用如下回归：

```
gen log_income=log(labor_inc)
reg log_income i.a2012
```

而注意到由于抽样的问题，每个样本的重要性是不一样的，我们可以使用swgt变量作为权重进行加权：

```
gen log_income=log(labor_inc)
reg log_income i.a2012 [w=swgt]
```

得到的结果与不加权的相比有一些变化，两个结果相比较，不加权的回归系数可能低估了高教育程度与低教育程度的收入差距。

加总数据的权重

- 正如在加总数据中计算均值可能需要加权平均一样，在一些加总数据中，也需要使用加权最小二乘。
- 比如，如果对于每一个个体，有回归方程：

$$y_{ig} = x'_{ig}\beta + u_{ig}, g = 1, ..., G$$

其中g为一个分组变量，比如省份或者城市

- 但是我们观察不到个体数据，只能使用平均数据：

$$\bar{y}_g = \bar{x}'_g\beta + \bar{u}_g, g = 1, ..., G$$

此时我们可以认为g组的N_g个个体被一个抽象的“平均个体”$(\bar{y}_g, \bar{x}'_g)'$所代表了，因而类比于逆概率加权，可以使用N_g作为权重进行加权。

加总数据中的异方差

在上面的例子中使用加权最小二乘同时可能还处理了异方差问题。

- 如果假设$u_{ig} \sim (0, \sigma^2)$，那么$\bar{u}_g \sim (0, \sigma^2/N_g)$，从而出现了异方差问题：方差随着每个城市人口的变化而变化。
- 此时，我们可以考虑在方程两边同时乘以$\sqrt{N_g}$，得到：

$$\sqrt{N_g}\bar{y}_g = \sqrt{N_g}\bar{x}'_g\beta + \sqrt{N_g}\bar{u}_g$$

那么此时$\sqrt{N_g}\bar{u}_g \sim (0, 1)$，从而消去了异方差问题。

- 再进行最小二乘回归，就得到了：

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^{N}\left[\left(\sqrt{N_g}\bar{x}_g\right)\left(\sqrt{N_g}\bar{x}_g\right)'\right]\right)^{-1}\left(\sum_{i=1}^{N}\left[\left(\sqrt{N_g}\bar{x}_g\right)\left(\sqrt{N_g}\bar{y}_g\right)\right]\right) \\ &= \left(\sum_{i=1}^{N}\left[N_g\bar{x}_g\bar{x}'_g\right]\right)^{-1}\left(\sum_{i=1}^{N}\left[N_g\bar{x}_g\bar{y}'_g\right]\right)\end{aligned}$$

上式无非就是使用N_g作为权重的加权最小二乘。

加权最小二乘

美国单边离婚法案

Friedberg（1998）在研究单边离婚法案（unilateral divorce law）对美国离婚率的影响时，使用了如下设定：

$$divrate = b_0 + b_1 \times unilateral + x'\beta + u$$

由于离婚率可以看成是一个州每个女性是否离婚的均值，所以他们在回归时使用州的人口作为权重，以下代码展示了他们的基础回归结果：

加权最小二乘

美国单边离婚法案

```
clear
set more off
use "datasets/Divorce-Wolfers-AER.dta"
egen state=group(st)
// reg
reg div_rate unilateral divx* i.state i.year if year>1967 & year<1989
reg div_rate unilateral divx* i.state i.year if year>1967 & year<1989 [w=stpop]
```