

拟合的其他方法

慧航

2023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

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

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

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：

```

1 use datasets/citydata.dta , clear
2 keep if year==2010
3 su v210
4 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：

```
1 gen log_income=log(labor_inc)
2 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：

```
1 gen log_income=log(labor_inc)
2 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：

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

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：

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

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：

核函数

权重一般可以如下选取：

- 令 $K(x)$ 为一个以纵轴对称的函数，即 $K(x) = K(-x)$ ，从而 $\int_{\mathbb{R}} xK(x) dx = 0$ ，且在 $x \in [0, \infty)$ 是单调递减的， $K(x) \geq 0$
- 令权重：

$$w_i = K\left(\frac{x_i - x_0}{h}\right)$$

其中 $h > 0$ 为窗宽 (bandwidth)，而 $K(x)$ 被称为“核函数” (kernel function)

- 一般核函数可以选取为对称分布的密度函数。

Rectangle核函数

如果令：

$$K_0(x) = \begin{cases} 1 & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

那么权重为：

$$w_i = K_0\left(\frac{x_i - x_0}{h}\right) = \begin{cases} 1 & x_0 - h \leq x_i \leq x_0 + h \\ 0 & otherwise \end{cases}$$

带入到目标函数中去，就得到了：

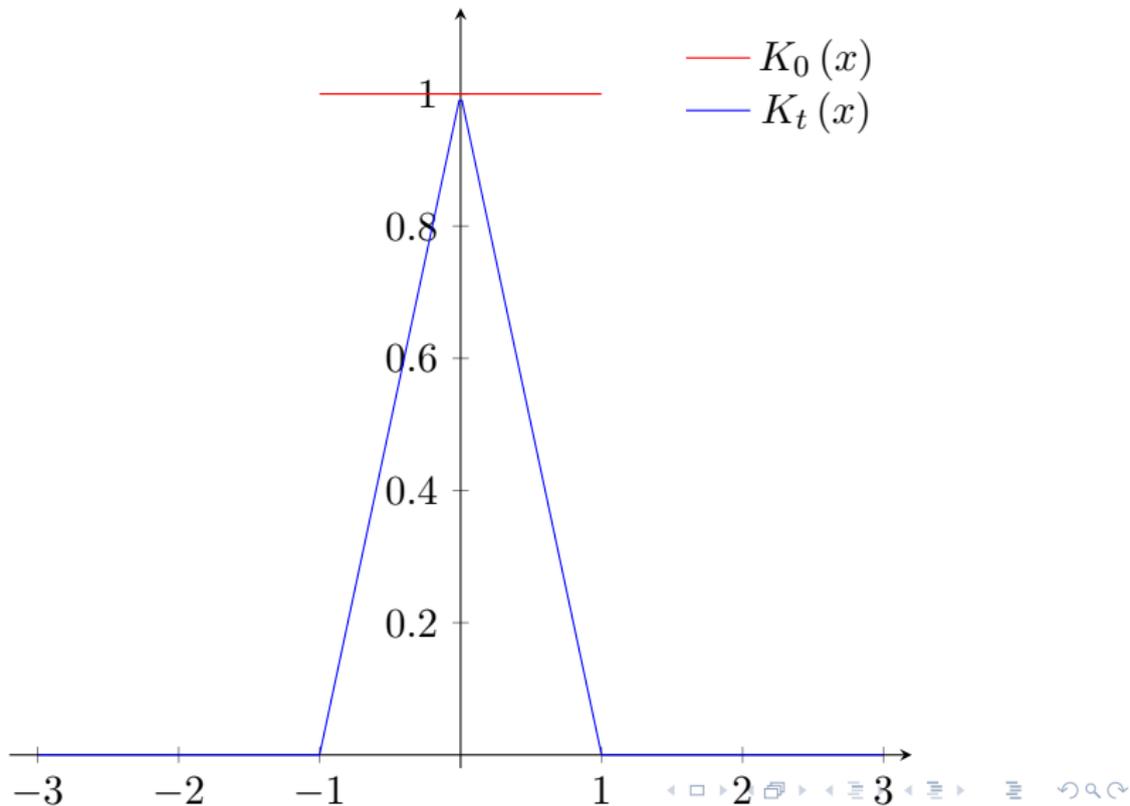
$$\min_{\beta} \sum_{i=1}^N \left[1_{\{|x_i - x_0| \leq h\}} (y_i - \alpha - \beta x_i)^2 \right]$$

以上目标函数等价于在区间 $[x_0 - h, x_0 + h]$ 内做简单的最小二乘法。

Rectangle核函数

- 这也就是 h 取名“窗宽”的由来：
 - h 决定了 x_0 附近区间的大小， h 越小，则在 x_0 附近取的区间越小，使用的数据量也就越少。
- 注意到 $K_0(x)$ 在临近的区间内权重都相等，所以我们也称其为矩形 (rectangle) 核函数

核函数



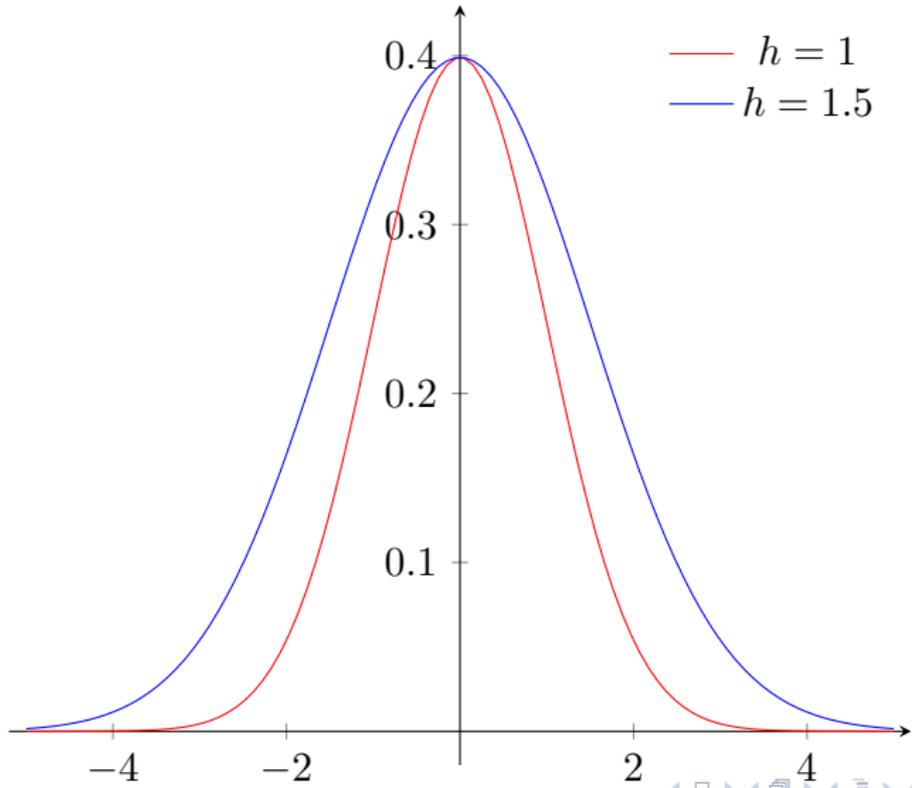
Triangle核函数

- 我们还可以令权重在临近区间内不相等：当 $x_{\{i\}}$ 距离 $x_{\{0\}}$ 越近时权重越高
- 此时可以在 $K_0(x)$ 的基础上进行修改，使用：

$$K_t(x) = \begin{cases} 1 - |x| & |x| \leq 1 \\ 0 & |x| > 1 \end{cases}$$

- 该函数同样在 $|x| > 1$ 时取值为0，但是在 $|x| \leq 1$ 这个区间内，随着 x 远离0，有个递减的过程。
- 我们称该核函数为三角（triangle）核函数。

核函数



局部常数估计

一个模拟

假设数据生成过程：

$$y = \exp(\sin x^3) + u$$

其中 $x \sim U(0, 2)$, $u \sim N(0, 1)$ ，假设样本量 $N = 300$ 。如果我们关心的是当 $x = 2$ 时 y 的预测值（真实值为 2.6895），我们选取标准正态分布的密度函数作为核函数，此外选取 $h = 0.1$ ，计算得到 $\hat{y}_{x=2} = 2.07$ (local_constant.do)。

局部线性估计

- 注意到，在上例中，局部常数项回归计算的实际上是在 $x = 2$ 的一个小的邻域中的均值
- 然而在上例中，由于 $x=2$ 恰好是 x 的取值范围的上界，所以我们实际上只使用了 $x=2$ 左边的一个小的邻域 $(2 - h, 2)$ 。
- 可以想象，由于在 $x = 2$ 左边，真实的数据生成过程是单调递增的，所以在这个小邻域中计算均值会低估 $\mathbb{E}(y|x = 2)$ 。
 - 上例的结果也可以看出，局部常数项的估计的确低估了 $\mathbb{E}(y|x = 2)$ 。

窗宽和多项式阶数选取

- 多项式阶数并不是越多越好，过高的多项式阶数会导致预测结果，特别是在端点的预测效果不稳定。
- 而关于窗宽 h ，考虑局部常数项回归以及 $K(\cdot) = K_0(\cdot)$ 作为例子，此时的估计量无非是 $(x_0 - h, x_0 + h)$ 区间的所有 y_i 的平均数
 - 可以想象一个过小的窗宽意味着能够使用的样本量更少，所以估计量的方差会很大；但是由于窗口比较小，我们上面所讨论的“低估”就会更不明显，也就是说估计量的偏差 (bias) 会更小。
 - 而反过来，一个大的窗宽会降低估计量的方差，但是偏差则会提高。
- 回忆均方误差可以写为偏差的平方和方差之和，所以理论上应该会有一个最优的 h ，使得均方误差达到最小。

选取方法

一般多项式阶数和窗宽的选取方法：

- 理论推导计算
 - 一些特殊情况有理论计算结果，如非参数回归、RD设计等
- 交叉验证
 - 我们仅仅关注 $x = x_0$ 处的预测，所以在做交叉验证时，并不需要所有的样本点都作为测试集，而是仅仅把离 x_0 最近的一些点作为测试集就好了。

窗宽和多项式阶数选取

交叉验证选取多项式阶数和窗宽

接上例，我们使用local_poly_cv.do代码，在 $p = 1, 2, \dots, 5$ 阶多项式、 $h = 0.01, 0.02, \dots, 0.5$ 的范围内搜索最优的 p 和 h 的组合。在以上代码中，我们针对每一个 (p, h) 的组合，都使用与 $x = 2$ 最近的10个点作为测试集，用留一验证的方法在测试集上计算交叉验证的均方误差，最后选取交叉验证均方误差最小的 (p, h) 的组合，并进行了局部多项式的回归。选取的结果是当 $h = 0.48, p = 3$ 时，均方误差最小，此时预测为 $\hat{y}_{x=2} = 2.743$ ，与真实值2.6895非常接近。

多个解释变量

- 现在如果我们有两个解释变量，即 $x = (x_1, x_2)'$ ，如果需要预测 $\mathbb{E}(y|x_1 = x_1^0, x_2 = x_2^0)$ ，那么可以在 x 的两个维度上分别取一个核函数和一个窗宽，并使用两个变量核函数的乘积作为权重：

$$w_i = K_1 \left(\frac{x_{1i} - x_1^0}{h_1} \right) K_2 \left(\frac{x_{2i} - x_2^0}{h_2} \right)$$

然后做加权最小二乘就可以了。

- 由于必须在多个维度都很接近 x_0 才会对 $\mathbb{E}(y|x = x_0)$ 的估计有贡献，然而在有限样本的情况下，随着维度的增加， x_0 附近的点会越来越少，为了保证相对较小的方差，就必须扩大范围，而扩大范围会造成比较大的偏差，所以在多维解释变量的情况下，以上的局部多项式估计可能并不理想，我们将这种现象称为“维数的诅咒”（the curse of dimensionality）。