体育经济分析: 原理与应用

单元4: 体育与计量经济

周正卿

20 February 2023

大纲

大纲

- 简要历史
- 经济效益

条件期望回顾

概率分布函数与概率密度函数

- 感兴趣的变量是工资
  - 它是个随机变量
  - 假设总体有确定的分布

Q: 要想了解总体特征，该如何实现?

- 可以抽样获得一个样本，用样本特征描述总体特征
  - 用样本分布估计总体分布，但两者之间是有“距离”的。之后解决。
  - 数学上的结论：概率分布函数（probability distribution function）可微，那么它的概率密度函数（probability density function）就能够反映概率分布函数的特点。
  - 可以制作频率直方图代表样本分布，来反映总体特征。（想象：将总体工资分成小区间，将抽到的值放入对应的区间，工资在每个区间内出现的次数）. 数学上 $f(w) = \frac{d}{dw}F(w)$，w 表示 $wage$

条件分布 (Conditional Distribution)

- 可以捕捉两个变量的关系。
- 假设 Y 与 X 是随机变项（量）.
  - Y 是因变量（被解释变量 | 结果变量）; X 是自变量（解释变量 | 干预变量）.
  - 随机变量（r.v.），具有概率分布
- 可以建立联合概率分布函数（joint probability distribution function）和联合密度函数 (joint density function) 来捕捉两个变量的关系。

条件期望（Conditional Expectation ）

- 概述
  - 更关心 X 和 Y 间的关系
  - 条件期望是描述这种关系的一种方法

工资（Y）和性别（X）差异的关系?

- 《E》p16:

工资对数的条件均值可以写成如下形式:

$$E[log(wage) \mid gender = man] = 3.05$$

$$E[log(wage) \mid gender = woman] = 2.81$$

关注条件均值的好处：将复杂分布的特点描述简单（均值），方便组间比较。 条件均值是经济分析和回归分析的主要关注点。还可以增加其他的条件，种族，后的工资比较。

$$E[\log(wage) \mid gender = \text{man}, race = white] = 3.07$$

$$E[\log(wage) \mid gender = woman, race = black] = 2.73$$

条件期望函数（Conditional Expectation Function）

当涉及多个“条件”时，可以写作：

$$E\left[Y \mid X_1=x_1, X_2=x_2, \ldots, X_k=x_k\right]=m\left(x_1, x_2, \ldots, x_k\right)$$

向量形式：

$$E[Y \mid X=x]=m(x)$$

所以 CEF $m(x)=E[Y \mid X=x]$ 就是 $x \in \mathbb{R}^k$的函数，意味着“ 当 X 取值 x 时, Y 的平均值为 $m(x)$ ”,由于 X 可以取值任意的 x , 因此将CEF视为随机变量 X 的函数。

Key: 深刻理解条件期望函数是 x 的函数

例：三个种族， $x=$(黑，白，其他) , $y=log(wage)$ ，每个种族都有一个工资的均值，均值与种族取值(x)一一对应关系。

边缘密度函数与条件密度函数

给定联合密度函数 $f(y,x)$, 变量 x 的边缘密度函数为:

$$f_X(x)=\int_{-\infty}^{\infty} f(y,x)dy$$

对于任意 x 的 $f_X(x)>0$, 给定 X , Y 的条件密度函数为:

$$f_{Y|X}(y\mid x)=\frac{f(y,x)}{f_X(x)}$$

条件密度相当于联合密度 $f(y,x)$ 在保持x不变情况下的随机化“切片”.

条件密度函数

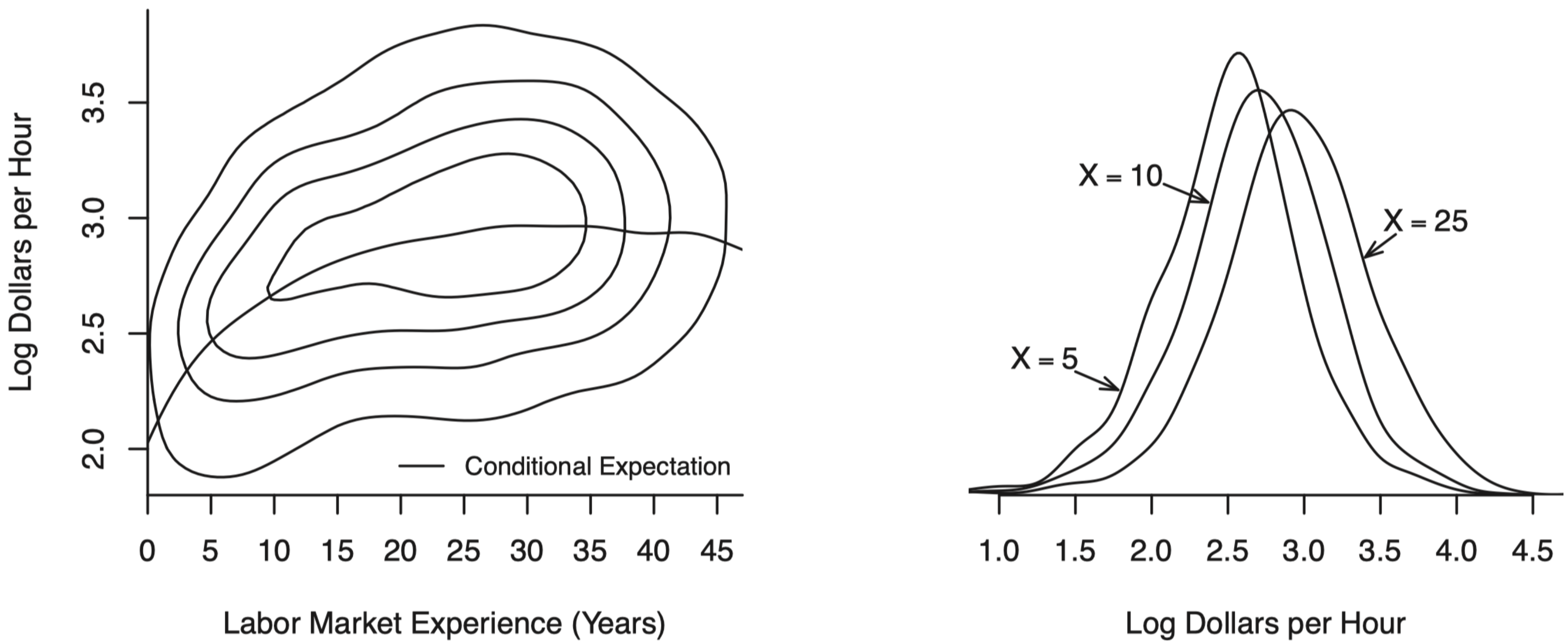
- 离散形式：

$$P(y|x) = \frac{P(y,x)}{P(x)}$$

其中 $P(x) = \sum_{i=1}^{N} P(y_i, x)$

边缘密度函数与条件密度函数

Q: 想象一下?



(a) Joint Density of Log Wage and Experience (b) Conditional Density of Log Wage given Experience

Figure 2.4: Log Wage and Experience

条件期望值函数的性质

- **性质1** (期望迭代法则,law of iterated expectation)

$$E[E[Y \mid X]] = E[Y]$$

$E[Y|X]$ 的期望值是 $[Y]$ 的无条件期望值。

例如：

$$\begin{aligned}
&\mathbb{E}\left[\log(wage) \mid gender = man\right] \mathbb{P}\left[gender = man\right] \\
&+ \mathbb{E}\left[\log(wage) \mid gender = woman\right] \mathbb{P}\left[gender = woman\right] \\
&= \mathbb{E}\left[\log(wage)\right].
\end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.81 \times 0.43 = 2.95.$$

- **性质1**推论

$$E\left[E[Y|X_1,X_2]|X_1\right]=E[Y|X_1]$$

  - 内部期望值以X1和X2同时为条件,外部期望值只以X1为条件。迭代后的期望值可以得到简单的答案E[Y|X1],即只以X1为条件的期望值。《E》表述为"较小的信息集获胜" .mono[-->] 以小谋大

例：

$$\begin{aligned}
&\mathbb{E}\left[\log(wage)\mid gender=man,\ race=white\right]\mathbb{P}\left[race=white\mid gender=man\right]\\
&+\mathbb{E}\left[\log(wage)\mid gender=man,\ race=Black\right]\mathbb{P}\left[race=Black\mid gender=man\right]\\
&+\mathbb{E}\left[\log(wage)\mid gender=man,\ race=other\right]\mathbb{P}\left[race=other\mid gender=man\right]\\
&=\mathbb{E}\left[\log(wage)\mid gender=man\right]
\end{aligned}$$

or numerically

$$3.07\times 0.84+2.86\times 0.08+3.03\times 0.08=3.05.$$

- **性质2** (线性)

$$E[a(X)Y + b(X)|X] = a(X)E[Y|X] + b(X)$$

对于函数 $a(\cdot)$ and $b(\cdot)$.

- **性质3**（独立意味着均值独立）

若 X 与 Y 独立, 则 $E[Y|X] = E[Y]$

- 性质3证明 (以离散变量为例):

$$
\begin{aligned}
E[Y|X] &= \sum_{i=1}^{N} y_i P(Y = y_i|X) \\
&= \sum_{i=1}^{N} y_i \frac{P(Y = y_i, X)}{P(X)} \\
&= \sum_{i=1}^{N} y_i \frac{P(Y = y_i) \times P(X)}{P(X)} \quad = E[Y].
\end{aligned}
$$

用到 $P(Y = y, X = x) = P(X = x)P(Y = y)$.

- **性质4** （均值独立意味着不相干）

  若 $E[Y|X] = E[Y]$, 则 $Cov(X, Y) = 0$.
  - $E[Y|X] = E[Y]$ is 均值独立(**mean independence**)
  - 记住: 均值独立意味着不相干，反过来不一定成立.

- **性质5** （条件期望值是最小均值平方误差）

假设对于任意函数 g 有 $E[Y^2] < \infty$ 并 $E[g(X)] < \infty$, 那么

$$E[(Y - \mu(X))^2] \leq E[(Y - g(X))^2]$$

其中 $\mu(X) = E[Y|X]$.

- 解读:
  - 假设使用某种函数形式 g 和数据 X 来解释 Y
  - 那么 g 的最小均方误（ **the mean squared error** ）就是条件期望。

- **性质5** 证明(自行推导):

$$
\begin{aligned}
E[(Y-g(X))^2] &= E[\{(Y-\mu(X))+(\mu(X)-g(X))\}^2] \\
&= E\left[(Y-\mu(X))^2\right]+E\left[(\mu(X)-g(X))^2\right] \\
&+ 2E\left[(Y-\mu(X))(\mu(X)-g(X))\right].
\end{aligned}
$$

使用期望迭代法则

$$
\begin{aligned}
E\left[(Y-\mu(X))(\mu(X)-g(X))\right] &= E\left\{E\left[(Y-\mu(X))(\mu(X)-g(X))\,|X\right]\right\} \\
&= E\left\{(\mu(X)-g(X))\left(E[Y|X]-\mu(X)\right)\right\} \\
&= 0
\end{aligned}
$$

所有,

$$
E[(Y-g(X))^2] = E\left[(Y-\mu(X))^2\right]+E\left[(\mu(X)-g(X))^2\right],
$$

which can take its minimum when $g(X)=\mu(X)$.

其他有用的性质

- **概率迭代法则**

$$P(Y) = \sum_{i=1}^{N} P(Y|x_i)P(x_i)$$

  X是离散随机变量

- **方差加法法则**

$$Var(Y) = E[V(Y|X)] + V[E(Y|X)]$$

误差与模型构建

条件期望函数误差

- Conditional Expectation Function Error （CEFE）

$$e = Y - E(Y|X) = Y - m(x)$$

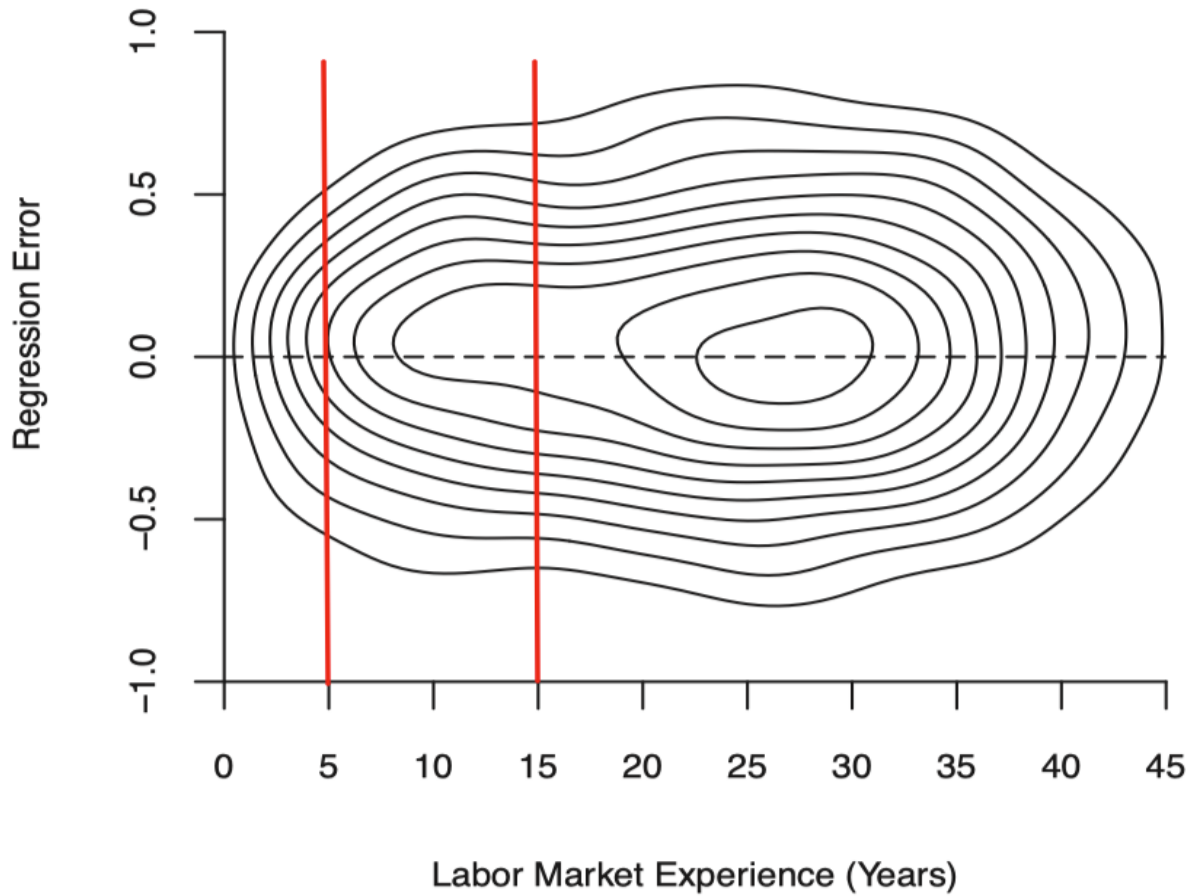
  - X 是RVs, $E(Y|X)$ 也是RVs.对于二元变量 D_i , CEF有两个值 $E[Yi|Di = 1]$和 $E[Y_i|D_i = 0]$
  - e 是RVs，具有概率分布

- CEFE性质

  1. $E(e|X) = 0$
  2. $E(e) = 0$
  3. 对于任意形式 $h(x)$,
     $E(h(X) \cdot e) = 0$

条件期望函数误差（图示）

Key: 注意条件分布的形状是随着工作经验如何变化?



总结：模型构建（by Hansen）

step1: 定义条件期望函数 $m(x) = E(Y|X)$

step2: 定义条件期望函数误差 $e = Y - m(x)$

推导出：

$$Y = m(x) + e$$

因此模型类别由 $m(x)$ 形式决定。

如截距模型，线性模型，Logit模型等。