分位数回归

相耐汀

2019.11.5

主要内容

- 为什么需要分位数回归
- 7.1分位数回归模型
  - 7.1.1删失分位数回归
  - 7.1.2分位数回归的近似性质
  - 7.1.3微妙之处
- 7.2对分位数处理效应的工具变量估计
  - 7.2.1分位数处理效应估计值

为什么需要分位数回归

为什么需要分位数回归

- 分位数回归是给定回归变量X，估计响应变量Y条件分位数的一个基本方法。
- 分位回归优点：
  - 能够更全面的描述被解释变量条件分布的全貌，而不仅仅分析被解释变量的均值。
  - 中位数回归与最小二乘法相比，估计结果对离群值表现得更加稳健。
- 举例：工资分布的变化

7.1分位数回归模型

条件分位数函数

- 假定连续分布随机变量Y_i密度函数性质良好，那么给定回归元所在向量X_i，在分位数τ处的条件分位数函数可定义为：

$$Q_\tau(Y_i|X_i) = F_y^{-1}(\tau|X_i)$$

- 其中$F_y(y|X_i)$是在给定X_i时Y_i在Y处的分布函数。

条件分位数函数

- 条件分布函数实际上是一类特殊的条件期望函数
- 条件期望函数：

$$E[Y_i|X_i] = \underset{m(X_i)}{argmin}\, E[(Y_i - m(X_i))^2]$$

条件期望函数就是给定X_i后，对Y_i的最小均方误预测

- 条件分位函数也是最小化问题的解：

$$Q_\tau(Y_i|X_i) = \underset{q(X_i)}{argmin}\, E[\rho_\tau(Y_i - q(X_i))]$$

其中$\rho_\tau(u) = (\tau - 1(\mathrm{u} \le 0))u$被称为“校验函数”

条件分位数函数

- 当τ=0.5时
  - $\rho_{0.5}(u) = \frac{1}{2}(sign\ u)u = \frac{1}{2}|u|$
  - 最小化问题转化为离差绝对值最小值

$$Q_\tau(Y_i|X_i) = \underset{q(X_i)}{argmin}\, E[\frac{1}{2}|Y_i - q(X_i)|]$$

- 其他情况下
  - $\rho_{\tau(u)} = 1(u > 0) * \tau|u| + 1(u \leq 0) * (1-\tau)|u|$

即

$$\rho_{\tau(u)} = \begin{cases} (\tau - 1)u, & u \leq 0 \\ \tau u, & u > 0 \end{cases}$$

  - 通过不对称加权可以产生一个最小化元，从而计算出条件分位数

分位数回归

- 分位数回归模型

$$Q_\tau(Y_i|X_i) = X_i'\beta_\tau$$

其中

$$\beta_\tau = \underset{b}{argmin}\, E[\rho_\tau(Y_i - X_i'b)]$$

教育水平对工资分布影响

- 以工资分布的研究为例来探讨分位数回归的使用
- 劳动经济学家感兴趣的问题：给定类似教育水平和工作经验等协变量后，工资不平等是如何变化的。
- 在20世纪80和90年代，在不同教育水平组成的群体之间的总体收入差距有显著提升（比如接受高等教育的工资溢价），但是我们并不清楚在具有相同教育水平和工作经验的群体内部，工资水平是如何变化的。
- 以下用分位数回归的方法证明近年来群体内部工资不平等的扩大。

教育水平对工资分布影响

Table 7.1.1: Quantile regression coefficients for schooling in the 1970, 1980, and 2000 Censuses

		Desc. Stats.		Quantile Regression Estimates					OLS Estimates	
Census	Obs.	Mean	SD	0.1	0.25	0.5	0.75	0.9	Coeff.	Root MSE
1980	65023	6.4	0.67	.074 (.002)	.074 (.001)	.068 (.001)	.070 (.001)	.079 (.001)	.072 (.001)	0.63
1990	86785	6.46	0.06	.112 (.003)	.110 (.001)	.106 (.001)	.111 (.001)	.137 (.003)	.114 (.001)	0.64
2000	97397	6.5	0.75	.092 (.002)	.105 (.001)	.111 (.001)	.120 (.001)	.157 (.004)	.114 (.001)	0.69

Notes: Adapted from Angrist, Chernozhukov, and Fernandez-Val (2006). The tables reports quantile regression estimates of the returns to schooling, with OLS estimates shown at the right for comparison. The sample includes US-born white and black men aged 40-49. Standard errors are reported in parentheses. All models control for race and potential experience. Sampling weights were used for the 2000 Census estimates.

教育水平对工资分布影响

- 如果教育水平对工资的影响类似以一种位移
  - 用经典线性回归模型来描述

$$Y_i \sim N(X_i'\beta, \sigma_\varepsilon^2)$$

  - 则如果不考虑变化的截距项，在每个分位数水平上，分位数回归系数应相同

$$P[Y_i - X_i'\beta < \sigma_\varepsilon \Phi^{-1}(\tau)|X_i] = \tau$$

- 如果教育水平对工资的影响不是位移
  - 用经典线性回归模型来描述

$$Y_i \sim N(X_i'\beta, \sigma^2(X_i))$$

  - 令$\sigma^2(X_i) = (\lambda' X_i)^2$

$$P[Y_i - X_i'\beta < (\lambda' X_i)^2 \Phi^{-1}(\tau)|X_i] = \tau$$

则在各个分位数水平上$\beta_\tau = \beta + \lambda\Phi^{-1}(\tau)$分位数回归系数变化

7.1.1删失分位数回归

删失数据定义

- 分位数回归允许我们在Y_i的分布一段被隐藏时仍可以考察Y_i的条件分布特征。
- 假设所得数据形式如下

$$Y_{\mathrm{i},obs} = Y_i * 1[Y_i < c] + c * 1[Y_i \geq c]$$

其中$Y_{\mathrm{i},obs}$是看到的观测值，Y_i是本应该看到的数据。

变量$Y_{\mathrm{i},obs}$是Y_i的删失数据。

处理删失数据

- 由于我们不知道哪个条件分位数正处于删失点之下，所以（以高收入者收入被加密为例）考虑下式：

$$Q_\tau(Y_i|X_i) = \min(c, X_i'\beta_\tau^c)$$

- 参数β_τ^c为下式的解

$$\beta_\tau^c = \underset{b}{argmin}\, E\{1[X_i'b < c] \cdot \rho_\tau(Y_i - X_i'b)]$$

- 以上方法也就是针对$X_i'\beta_\tau^c < c$求解分位数回归的最小化问题，只要有足够多的未删失数据，我们求得的估计值就给出了原本在不存在删失数据时才能求得的分位数回归函数。

7.1.2分位数回归的近似性质

分位数回归的近似性质

- 我们可以将分位数回归理解为一种最小均方误意义下的对条件分位函数的线性近似。
- 定义分位数回归的设定偏误：

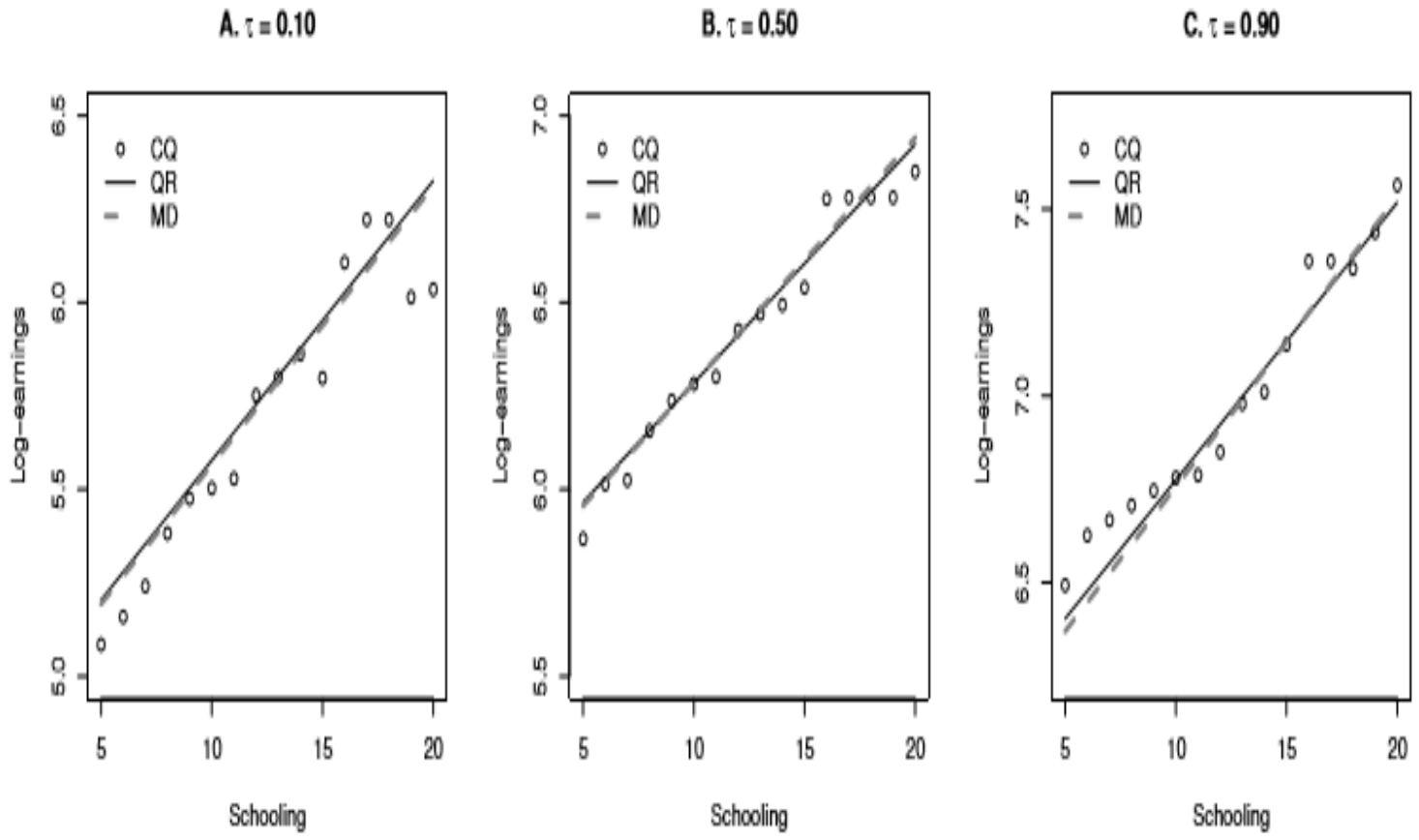
$$\Delta_\tau(X_i, \beta_\tau) = X_i'\beta - Q_\tau(Y_i|X_i)$$

- Angrist, Chernozhukov, and Fernandez-Val (2006) 指出：通过最小化模型设定偏误的加权平均值，我们可以得到总体分位数回归向量。

分位数回归的近似性质

- 可以用下述图片来阐明分位数回归的近似性质。
- 使用1980年的人口普查数据，给定最高完成学位水平下，绘出对数工资的条件分位数函数。
- 图A-C分别在0.1、0.5、0.9三个分位数水平下绘出了对$Q_\tau(Y_i|X_i)$的线性分位数回归结果。
- 图中圆圈表示针对不同教育水平的个体分别估计出的条件分位数函数；分位数回归曲线则是实线。

分位数回归的近似性质


A. τ = 0.10
B. τ = 0.50
C. τ = 0.90
CQ
QR
MD
Log−earnings
Schooling
5.0
5.5
6.0
6.5
7.0
7.5
5
10
15
20

7.1.3微妙之处

微妙之处

- 分位数回归的系数告诉我们的是对分布的影响，而不是对个体的影响。
- 条件分位数和边际分位数的转变。

求解边际分位数

- 假设条件分位数函数是线性的，即$Q_\tau(Y_i|X_i) = X_i'\beta_\tau$
- 令$F_y(y|X_i) = P[Y_i < y|X_i]$是在给定$X_i$时$Y_i$在Y处的条件累计密度函数,其边际分布为$F_y(y) = P[Y_i < y]$.
- 累计密度函数及其反函数关系为：

$$\int_0^1 1\left[F_y^{-1}(\tau|X_i) < y\right]d\tau = F_y(y|X_i)$$

- 用线性模型替换积分括号里面的条件分位函数，得

$$F_y(y|X_i) = \int_0^1 1[X_i'\beta_\tau < y]d\tau$$

- 用迭代期望率来求解边际分布函数$F_y(y)$

$$F_y(y|X_i) = E[\int_0^1 1[X_i'\beta_\tau < y]d\tau]$$

- 对函数$F_y(y)$进行转置，求得边际分位数$Q_\tau(Y_i)$

$$Q_\tau(Y_i) = \inf\{y: F_y(y) \geq \tau\}$$

7.2对分位数处理效应的工具变量估计

- 用工具变量法来解决遗漏变量偏误问题
- Abadie,Angrist,和 Imbens (2002)给出了使用工具变量估计分位数处理效应（QTE）的方法。
  - 该方法的假设和局部平均处理效应框架下的假设相同。
  - 对于$\tau \in (0,1)$,假设存在α_τ和β_τ满足:

$$Q_\tau(Y_i|X_i, D_i, D_{1i} > D_{0i}) = \alpha_\tau D_i + X_i'\beta_\tau$$

  - 则参数α_τ给出了给定X_i后依从工具变量者在Y_{1i}和Y_{0i}上的分位数之差。

$$\alpha_\tau = Q_\tau(Y_{1i}|X_i, D_{1i} > D_{0i}) - Q_\tau(Y_{0i}|X_i, D_{1i} > D_{0i})$$

7.2.1分位数处理效应估计值

分位数处理效应估计值

- 通过对依从工具变量者总体进行分位数回归，我们可以估计出依从工具变量者的分位数回归系数。
- $$(\alpha_\tau, \beta_\tau) = \underset{a,b}{argmin}\, E\{\rho_\tau(Y_i - aD_i + X_i'\mathrm{b}) | D_{1i} > D_{0i}\}$$
$$= \underset{a,b}{argmin}\, E\{k_i \rho_\tau(Y_i - aD_i + X_i'\mathrm{b})\}$$
- 其中，
$$k_i = 1 - \frac{D_i(1 - Z_i)}{1 - P(Z_i = 1|X_i)} - \frac{(1 - D_i)Z_i}{P(Z_i = 1|X_i)}$$

估计培训为受训者收入分位数影响

- 工作培训合法法案JTPA是一个大规模的联邦项目，他为20世纪80年代事业的美国工人提供资助性培训。
- 我们使用5102个成年按行来做培训对收入的影响的最小二乘回归、分位数回归、两阶段最小二乘回归和分位数处理效应估计值。
- JTPA实验中：
  - 控制组：没有人获得JTPA服务
  - 处理组：被分配到JTPA服务但只有60%的人接受。

估计培训为受训者收入分位数影响

A. OLS and Quantile Regression Estimates

	OLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
Training	3,754	1,187	2,510	4,420	4,678	4,806
	(536)	(205)	(356)	(651)	(937)	(1,055)
% Impact of Training	21.20	135.56	75.20	34.50	17.24	13.43
High school or GED	4,015	339	1,280	3,665	6,045	6,224
	(571)	(186)	(305)	(618)	(1,029)	(1,170)
Black	-2,354	-134	-500	-2,084	-3,576	-3,609
	(626)	(194)	(324)	(684)	(1087)	(1,331)
Hispanic	251	91	278	925	-877	-85
	(883)	(315)	(512)	(1,066)	(1,769)	(2,047)
Married	6,546	587	1,964	7,113	10,073	11,062
	(629)	(222)	(427)	(839)	(1,046)	(1,093)
Worked less than 13	-6,582	-1,090	-3,097	-7,610	-9,834	-9,951
weeks in past year	(566)	(190)	(339)	(665)	(1,000)	(1,099)
Constant	9,811	-216	365	6,110	14,874	21,527
	(1,541)	(468)	(765)	(1,403)	(2,134)	(3,896)

估计培训为受训者收入分位数影响

B. 2SLS and QTE Estimates

	2SLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
Training	1,593	121	702	1,544	3,131	3,378
	(895)	(475)	(670)	(1,073)	(1,376)	(1,811)
% Impact of Training	8.55	5.19	11.99	9.64	10.69	9.02
High school or GED	4,075	714	1,752	4,024	5,392	5,954
	(573)	(429)	(644)	(940)	(1,441)	(1,783)
Black	-2,349	-171	-377	-2,656	-4,182	-3,523
	(625)	(439)	(626)	(1,136)	(1,587)	(1,867)
Hispanic	335	328	1,476	1,499	379	1,023
	(888)	(757)	(1,128)	(1,390)	(2,294)	(2,427)
Married	6,647	1,564	3,190	7,683	9,509	10,185
	(627)	(596)	(865)	(1,202)	(1,430)	(1,525)
Worked less than 13	-6,575	-1,932	-4,195	-7,009	-9,289	-9,078
weeks in past year	(567)	(442)	(664)	(1,040)	(1,420)	(1,596)
Constant	10,641	-134	1,049	7,689	14,901	22,412
	(1,569)	(1,116)	(1,655)	(2,361)	(3,292)	(7,655)