

二元选择与广义线性模型

司继春

上海对外经贸大学

2026年9月



二元选择模型

二元选择模型用于解决被解释变量 y 只有两个取值的情况：

- 是否选择去上大学（上大学=1，不上大学=0）
- 某个国家是否爆发内战（爆发=1，不爆发=0）
- 企业是否选择出口（出口=1，不出口=0）
-

针对这类问题，由于被解释变量只有两种取值（二元变量），因而统称为『二元选择模型（binary choice model）』

二元选择模型

- 如果 y 的取值范围 $\text{supp}(y) = \{0, 1\}$ ，此时其条件期望函数：

$$\begin{aligned}\mathbb{E}(y|x) &= 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) \\ &= P(y = 1|x)\end{aligned}$$

即给定 x ， $y = 1$ 的概率

- 如果设定：

$$\mathbb{E}(y|x) = P(y = 1|x) = x'\beta$$

即线性概率模型（linear probability model），那么 $x'\beta$ 可能小于0或者大于1，对于预测而言就不太合适了。

二元选择模型

- 我们可以使用一个分布函数 $F(\cdot)$ 将 $x'\beta$ 压缩到 $[0, 1]$ 范围以内：

$$\mathbb{E}(y|x) = \mathbb{P}(y = 1|x) = F(x'\beta)$$

- 如果 $F(\cdot)$ 取标准的Logistic分布的分布函数，则称为Logistic回归或者Logit回归：

$$\mathbb{E}(y|x) = \mathbb{P}(y = 1|x, \beta) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} \triangleq \Lambda(x'\beta)$$

其中

$$\Lambda(x) = \frac{1}{1 + e^{-x}}$$

Logistic回归系数的解释

我们将 $y_i = 1$ 的概率与 $y_i = 0$ 的概率的比值成为几率（odds）：

$$\text{odds} = \frac{P(y = 1|x, \beta)}{P(y = 0|x, \beta)}$$

- 几率度量了 $y = 1$ 的概率相对于 $y = 0$ 的概率的大小
 - 如果几率等于1，意味着 $y = 1$ 的概率等于 $y = 0$ 的概率
 - 如果几率大于1，意味着 $y = 1$ 的概率大于 $y = 0$ 的概率（比如odds = 2代表概率为2倍）
 - 如果几率小于1，意味着 $y = 1$ 的概率小于 $y = 0$ 的概率（比如odds = 0.5代表概率为一半）

几率比

几率比 (odds ratio)：两个不同组的几率的比率，如果等于1则代表两个组的 $y = 1$ 的概率相同。

几率比

我们使用CFPS数据计算了是否退出劳动力市场与性别的频数，并计算男性和女性退出劳动力市场的odds：

	女性	男性
未退出	1289	1722
退出	255	42
Odds	0.1978	0.0244

从而，女性退出劳动力市场的odds为男性的 $0.1978/0.0244 = 8.1$ 倍，女性退出劳动力市场的概率更高。

Logistic回归

对于Logistic回归：

$$P(y = 1|x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}, P(y = 0|x) = \frac{1}{1 + e^{x'\beta}}$$

从而几率 (odds) 为：

$$\text{odds} = \frac{P(y = 1|x, \beta)}{P(y = 0|x, \beta)} = \frac{\frac{e^{x'\beta}}{1 + e^{x'\beta}}}{1 - \frac{e^{x'\beta}}{1 + e^{x'\beta}}} = e^{x'\beta}$$

从而对数几率 (log odds, 也称为logit) 为：

$$\text{logit} = \ln(\text{odds}) = x'\beta$$

因而以上模型被称为对数几率回归 (Logit regression) .

估计

- 估计：以上模型的条件密度函数为：

$$f(y_i|x_i, \beta) = \left[\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right]^{\mathbb{1}\{y_i=1\}} \left[\frac{1}{1 + e^{x_i'\beta}} \right]^{\mathbb{1}\{y_i=0\}}$$

因而对数似然函数为：

$$L(\beta|Y, X) = \sum_{i=1}^N \left[y_i \ln \left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right) + (1 - y_i) \ln \left(\frac{1}{1 + e^{x_i'\beta}} \right) \right]$$

最大化以上对数似然函数即可。

- 预测：

$$\widehat{\mathbb{E}(y|x)} = \hat{p}(x) = F(x'\hat{\beta})$$

Probit模型

- 如果 $F(\cdot)$ 取标准正态分布的分布函数，则称为Probit回归：

$$\mathbb{E}(y|x) = P(y = 1|x) = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

其对数似然函数为：

$$L(\beta|Y, X) = \sum_{i=1}^N [y_i \ln(\Phi(x'_i\beta)) + (1 - y_i) \ln(1 - \Phi(x'_i\beta))]$$

Probit、Logit回归的F.O.C.

论是Logit回归还是Probit回归，其对数似然函数都可以写为：

$$L(\beta|Y, X) = \sum_{i=1}^N [y_i \ln(F(x'_i\beta)) + (1 - y_i) \ln(1 - F(x'_i\beta))]$$

其一阶条件为：

$$\begin{aligned} \frac{\partial L(\beta|Y, X)}{\partial \beta} &= \sum_{i=1}^N \left[y_i \frac{1}{F(x'_i\beta)} f(x'_i\beta) x_i - (1 - y_i) \frac{1}{1 - F(x'_i\beta)} f(x'_i\beta) x_i \right] \\ &= \sum_{i=1}^N \left[\frac{y_i}{F(x'_i\beta)} - \frac{(1 - y_i)}{1 - F(x'_i\beta)} \right] f(x'_i\beta) x_i \\ &= \sum_{i=1}^N \frac{f(x'_i\beta)}{F(x'_i\beta) [1 - F(x'_i\beta)]} [y_i - F(x'_i\beta)] x_i = 0 \end{aligned}$$

Pseudo- R^2

类比于线性回归中的 R^2 ，McFadden (1974) 建议使用“伪 R^2 ” (pseudo- R^2)：

$$\begin{aligned} R^2 &= 1 - \frac{L(\hat{\beta}|Y, X)}{L_0} \\ &= 1 - \frac{\sum_{i=1}^N [\mathbb{1}\{y_i = 1\} \ln(\hat{p}_i) + \mathbb{1}\{y_i = 0\} \ln(1 - \hat{p}_i)]}{N [\bar{y} \ln \bar{y} + (1 - \bar{y}) \ln(1 - \bar{y})]} \end{aligned}$$

作为拟合优度的度量。

评价分类的指标

- 对于所有分类任务：选定一个临界值 c ，并令预测值 $\hat{y}_i = \mathbb{1} \{ \hat{p}_i > c \}$ ，之后将样本分为四类：
 - 真正 (True Positive, TP) : $y_i = 1, \hat{y}_i = 1$;
 - 假正 (False Positive, FP) : $y_i = 0, \hat{y}_i = 1$;
 - 真反 (True Negative, TN) : $y_i = 0, \hat{y}_i = 0$;
 - 假反 (False Negative, FN) : $y_i = 1, \hat{y}_i = 0$.

评价分类的指标

使用以上四个分类分别定义：

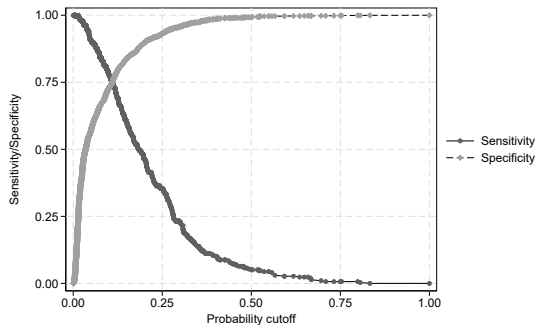
- 查准率 (**precision**)，即所有预测为正的样本中，正确的比例： $\text{Precision} = \frac{TP}{TP+FP}$ ；
- 查全率 (或者召回率, **recall**)，即所有正的样本中，正确的比例： $\text{Recall} = \frac{TP}{TP+FN}$ ；
- 精度 (**accuracy**)，即所有样本中预测正确的比例： $\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$ 。
- F1度量，即查准率和查全率的调和平均：

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{1}{\frac{1}{2} \left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \right)}$$

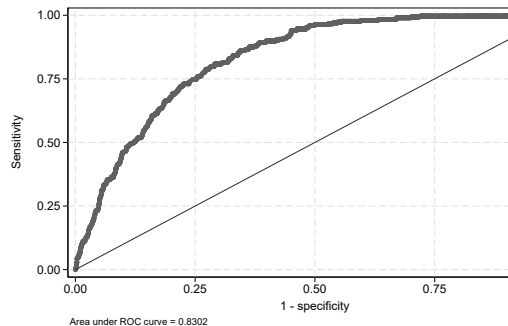
查准率和查全率之间通常存在着权衡：

- 比如，如果我们希望提高查准率，需减少预测为正的比例，需要较大的 c ，从而降低查全率。
- 使得查准率等于查全率的点称为平衡点 (break-event point, BEP)。

ROC曲线



(a) 敏感性曲线、特异性曲线



(b) ROC曲线

(logit_and_roc.do)

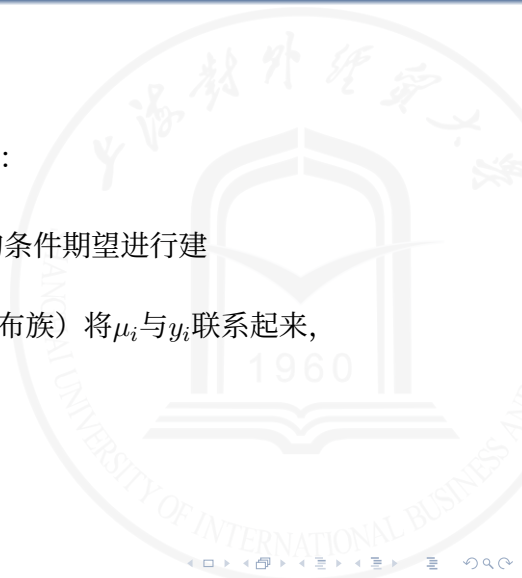
广义线性模型

- 为了拟合条件期望函数，以上针对不同的因变量的数据形式，我们分别使用了：线性回归、Probit回归以及Logit回归
- 以上回归都是广义线性模型（**generalized linear model**）的特例
- 广义线性模型是估计条件期望的一个比较一般的方法，特别是针对不同形式的因变量：
 - 分类变量
 - 计数变量
 - 受限的support
 -

广义线性模型

为了构建广义线性模型，需要以下几个组成部分：

- ① 首先将自变量 x 进行线性组合： $\eta_i = x_i' \beta$
- ② 使用一个链接函数（link function）对 y_i 的条件期望进行建模： $\mu_i = \mathbb{E}(y_i | x_i) = \mu(x_i' \beta)$
- ③ 使用一个期望为 μ_i 的分布 $\Psi(\cdot)$ （属于指数分布族）将 μ_i 与 y_i 联系起来，即 $y_i | x_i \sim \Psi(\mu_i)$.



广义线性模型

线性回归

如果我们假设链接函数为 $\mu_i = \mu(x_i'\beta) = x_i'\beta$ ，而选取

$$y_i|x_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

(假设 σ^2 已知)，我们知道正态分布属于指数分布族，因而以上就建立了一个广义线性模型. 注意到，对于以上模型，我们可以定义：

$$e_i = y_i - \mu_i = y_i - \mathbb{E}(y_i|x_i) \sim \mathcal{N}(0, \sigma^2)$$

或者：

$$y = x_i'\beta + e_i, e_i \sim \mathcal{N}(0, \sigma^2)$$

从而我们得到了普通线性回归模型.

广义线性模型：常用分布和链接函数组合

分布	支撑集	链接函数名	链接函数
正态分布	$\mathbb{R}$	identity	$\mu_i = x_i' \beta$
指数分布	$(0, +\infty)$	negative inverse	$\mu_i = -(x_i' \beta)^{-1}$
Gamma分布			
泊松分布	$\mathbb{Z}$	log	$\mu_i = \exp(x_i' \beta)$
负二项分布			
伯努利分布	$\{0, 1\}$	logit/probit	$\mu_i = N \cdot \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)}$
二项分布	$\{0, 1, \dots, N\}$		$\mu_i = N \cdot \Phi(x_i' \beta)$

广义线性模型：Stata

Logistic回归也可以使用如下广义线性模型的命令：

```
1 glm exit_labor `x', family(binomial) link(logit)
```

其中：

- family 为指数分布族
- link为链接函数

计数回归

泊松回归

如果被解释变量 y 是计数数据，即取值范围为自然数集 $\mathbb{Z}$ ，通常我们可以使用泊松分布对其进行建模：

$$y_i | \mu_i \sim \mathcal{P}(\mu_i)$$

其中 μ_i 为 $\mathbb{E}(y_i | x_i)$ 。考虑到 y_i 的取值范围为自然数，所以其期望不可能为负，为此在建模 μ_i 时，可以使用log链接函数，即： $\mu_i = \exp(x_i' \beta)$ 以上回归被称为泊松回归（Poisson regression）

计数回归

泊松回归

似然函数为

$$L(\beta|Y, X) = \sum_{i=1}^N (y_i x_i' \beta - \exp(x_i' \beta) - \ln y_i!)$$

一阶条件为

$$\frac{\partial L(\beta|Y, X)}{\partial \beta} = \sum_{i=1}^N [y_i - \exp(x_i' \beta)] x_i = 0$$

Poisson回归

OHIE数据

在数据集“OHIE_QJE.dta”中，记录了一些人参与某项健康保险计划之后的身体健康数据，其中 treatment 变量为是否参加该计划，而“baddays_phys_12m”为参与该计划12个月以后一个月内身体不舒服的天数，所以是一个计数数据。我们可以使用 glm 命令或者 poisson 命令对该回归进行计算：

```
1 use datasets/OHIE_QJE.dta
2 local x "treatment english_list female_list birthyear_list"
3 // 泊松回归
4 glm baddays_phys_12m `x', family(poisson) link(log)
5 // 与以下回归等价：
6 poisson baddays_phys_12m `x'
```

以上两条命令的结果是完全等价的。

计数回归

负二项回归

- 泊松分布虽然经常被用来建模计数问题，但是泊松分布的分布假设仍然太过于严格，比如，泊松分布的期望与方差相等，可能并不完全与现实情况吻合。
- 注意到如果 $X|\lambda \sim \mathcal{P}(\lambda)$, $\lambda \sim \Gamma\left(r, \frac{1-p}{p}\right)$, 那么 $X \sim \mathcal{NB}(r, p)$, 可以根据这一性质将泊松回归进一步拓展到负二项回归。

计数回归

负二项回归

- 如果假设

$$\mu_i = \exp(x_i' \beta + \nu_i)$$

$$e^{\nu_i} \sim \Gamma\left(\frac{1}{\alpha}, \alpha\right)$$

其中 $\alpha > 0$ 为过度离差 (overdispersion) 参数.

- 如果记 $\eta_i = e^{x_i' \beta}$, 假设 $\nu_i \perp x_i$, 根据伽马分布的定义,

$$\mu_i | x_i \sim \Gamma\left(\frac{1}{\alpha}, \alpha \eta_i\right)$$

从而

$$y_i | x_i \sim \mathcal{NB}\left(\frac{1}{\alpha}, \frac{1}{1 + \alpha \eta_i}\right)$$

非线性最小二乘

- 线性回归和广义线性模型中，我们都假设了条件期望函数 $\mathbb{E}(y|x)$ 为 x 的线性函数形式，注意这里的线性函数指的是参数线性，即 $\mathbb{E}(y|x) = x'\beta$ 为 β 的线性函数。
- 如果条件期望函数并非参数 β 的线性函数，此时我们可能需要使用非线性最小二乘法（nonlinear least squares, NLS）。

非线性最小二乘

Box-Cox

在上一节中我们介绍了Box-Cox变换. 如果我们假设

$$\mathbb{E}(y|x) = \beta_0 + \beta_1 \frac{x^\lambda - 1}{\lambda}$$

此时条件期望函数为参数 $\beta = [\beta_0, \beta_1, \lambda]'$ 的非线性函数, 就不能使用线性回归对以上方程进行估计了.

非线性最小二乘

- 一般地，我们可以设定条件期望为参数非线性的函数：

$$\mathbb{E}(y|x, \beta) = g(x, \beta)$$

从而定义误差项： $e = y - g(x, \beta)$

- 自然地，有 $\mathbb{E}(e|x) = 0$.
- 仿照线性回归的想法，我们可以最小化误差平方和，也就解最优化问题

$$\min_{\beta} \sum_{i=1}^N [y_i - g(x_i, \beta)]^2$$

从而得到参数 β 的估计.

非线性最小二乘

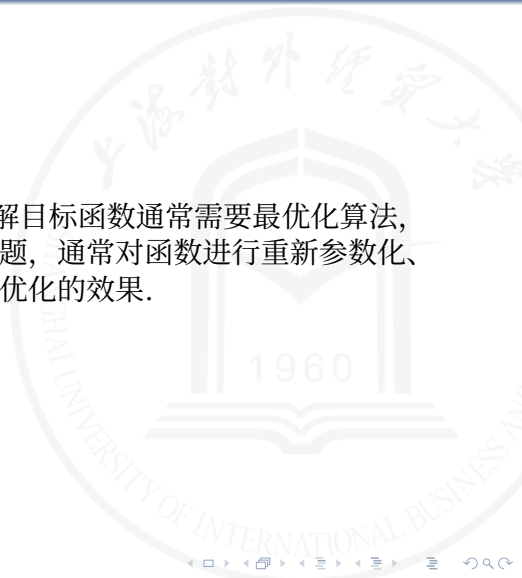
需要注意的是，在使用非线性最小二乘时，应该首先检查其识别条件，也就是不能存在两组参数： β_1 和 β_2 ，使得 $g(x, \beta_1) = g(x, \beta_2)$ 。

NLS的识别性

如果假设 $\mathbb{E}(y|x) = \beta \ln x^\gamma$ ，注意到经过简单变换以后，等价于 $\mathbb{E}(y|x) = \beta\gamma \ln x$ ，从而如果 $[\beta^*, \gamma^*]'$ 能够使得误差平方和最小，那么对于任意的 $c \neq 0$ ， $[c\beta^*, \gamma^*/c]'$ 也能够获得与 $[\beta^*, \gamma^*]'$ 相同的误差平方和，因而参数 β 和 γ 不能同时被识别。

非线性最小二乘

- 由于现在 $g(x, \beta)$ 为 β 的非线性函数，因而求解目标函数通常需要最优化算法，因而此时需要注意局部最优与全局最优的问题，通常对函数进行重新参数化、寻找比较好的初始值等都可能极大影响最优化的效果。



非线性最小二乘

边际消费倾向

为了估计边际消费倾向，一个简单的设定为：

$$C = \alpha + \beta Y + u$$

然而我们知道，边际消费倾向可能会随着收入水平的变化而变化，因而一个更一般化的设定是：

$$C = \alpha + \beta Y^\gamma + u$$

相应的边际消费倾向为：

$$\frac{\partial C}{\partial Y} = \beta \gamma Y^{\gamma-1}$$

当 $\gamma = 1$ 时，就是线性的函数；当 $\gamma > 1$ 时，意味着边际消费倾向会随着收入的增加而增加，反之则减少。（nls_mpc.do）

背景

- 有些变量虽然取值范围为正，但是仍然会有大量的0值存在，而这些变量往往分布是右偏的、带有增长性质的，取对数是一个常见的操作，但是因为0值的存在，取对数也会遇到困难。
- 为了解决这个问题，首先需要清楚大量0值存在的原因。
 - 这些0值有时可能会有经济行为上的解释，比如Tobit一类模型。
 - 除了这些经济行为上的原因，一个更为简单的原因可能仅仅是因为舍入误差（rounding error）的存在，即在数据记录时，那些非常贴近0但并非0的数字由于四舍五入的原因被直接记录为0。

舍入误差

舍入误差

- 如果

$$x \sim \mathcal{N}(1, 4)$$

$$u|x \sim \mathcal{N}(0, 0.1 \times e^{x-1})$$

$$y = \lfloor e^{x-1} + u + 0.5 \rfloor$$

其中 $\lfloor \cdot \rfloor$ 为向下取整，而 $\lfloor a + 0.5 \rfloor$ 即四舍五入。

- 以上数据生成过程，有大约37%的 y 取值为0。
- 以上数据生成过程中，我们假设了异方差的存在： $\mathbb{V}(u|x) = 0.1 \times e^{x-1}$ ，从而当 e^{x-1} 越贴近于0时，方差越小，反之则越大。

引力模型

引力模型的随机化

- 之前我们介绍了国际贸易中的引力模型，其中通过取对数的形式将连乘的形式变换为对数之和的形式. 如果加入误差项，那么取对数的版本为

$$\ln X_{ni} = \ln G + a \ln Y_i + b \ln Y_n + \ln \phi_{ni} + e_{ni}$$

- 然而，Silver和Tenreyro (2006) 认为， $X_{ni} = GY_i^a Y_n^b \phi_{ni}$ 这一关系应该在“平均”意义上成立，即 $GY_i^a Y_n^b \phi_{ni}$ 预测了贸易的均值，从而带有误差项的随机版本应该写为

$$X_{ni} = GY_i^a Y_n^b \phi_{ni} + e_{ni}$$

- 以上模型可以通过指数函数的形式进行改写，即

$$X_{ni} = \exp(\ln G + a \ln Y_i + b \ln Y_n + \ln \phi_{ni}) + e_{ni}$$

指数函数回归

- 以上两个例子中的模型都可以归结为如下模型：

$$y_i = \exp(x_i' \beta) + e_i$$

其中 $\mathbb{E}(e_i | x_i) = 0$.

- 针对以上模型，首先可以使用非线性最小二程进行估计，即解最优化问题

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

指数函数回归

- 一阶条件为

$$\sum_{i=1}^N ([y_i - \exp(x'_i \beta)] \cdot \exp(x'_i \beta) \cdot x_i) = 0$$

如果记 $\hat{e}_i = y_i - \exp(x'_i \hat{\beta})$ ，那么以上一阶条件可以理解为选取 $\hat{\beta}$ 使得 $\hat{e}_i$ 与 x_i 不相关，而 $\exp(x'_i \beta)$ 可以理解为权重。

- 从而，以上一阶条件赋予 $\exp(x'_i \beta)$ 更大的观测以更大的权重。

泊松伪极大似然

- 然而正如以上两个例子中所介绍的，以上模型可能存在严重的异方差问题。
- Silver和Tenreiro (2006) 假设条件方差 $\mathbb{V}(e_i|x_i) \propto \exp(x_i'\beta_0)$ ，从而

$$\mathbb{V}\left(\frac{e_i}{\sqrt{\exp(x_i'\beta_0)}} \middle| x_i\right) = \sigma^2$$

为常数

- 如果最小化加权版本的非线性最小二乘问题

$$\min_{\beta} \sum_{i=1}^N \left[\frac{y_i - \exp(x_i'\beta)}{\sqrt{\exp(x_i'\beta_0)}} \right]^2$$

泊松伪极大似然

- 一阶条件为

$$\sum_{i=1}^N \frac{[y_i - \exp(x'_i \beta)] \cdot \exp(x'_i \beta)}{\exp(x'_i \beta_0)} x_i = 0$$

- 考虑到当 $N \rightarrow \infty$ 时，将 $\beta = \beta_0$ 带入应该使得

$$\mathbb{E} \left(\frac{[y_i - \exp(x'_i \beta_0)] \cdot \exp(x'_i \beta_0)}{\exp(x'_i \beta_0)} x_i \right) = \mathbb{E} ([y_i - \exp(x'_i \beta_0)] x_i) = 0$$

成立，从而Silver和Tenreiro (2006) 建议直接解一阶条件

$$\sum_{i=1}^N ([y_i - \exp(x'_i \beta)] \cdot x_i) = 0$$

即可。以上一阶条件同样可以理解为选取 $\hat{\beta}$ 使得 $\hat{e}_i$ 与 x_i 不相关，但是将权重平等的赋予每个观测。

负二项与Gamma伪极大似然回归

- 类似的，Gourieroux、Monfort和Trognon（1984）提出如果同样假设 $E(y_i|x_i) = \exp(x_i'\beta)$ ，但是假设 y_i 的条件分布为负二项分布、 Γ 分布，那么就得到了负二项伪极大似然回归（NBPML）和Gamma伪极大似然回归（GPML）。
- Silver和Tenreiro（2006）使用数值模拟的方法比较了NLS、PPML以及GPML对于异方差以及舍入误差的稳健性，他们的结果发现其中PPML对于两者是相对稳健的。

不同方法的比较

Silver和Tenreyro (2006)

- 代码“log_of_gravity.do”使用Silver和Tenreyro (2006) 的数据复现了其论文中的表3的部分结果.
- Silver和Tenreyro (2006) 进一步使用Ramsey (1969) 的回归设定错误检验 (regression specification error test, RESET) 检验了每种方法所估计的条件期望函数的错误设定问题, 仍然发现在如上的引力方程模型中, PPML的结果没有发现模型错误设定.

分位数的定义

与之前类似，我们可以从总体和样本两个层面定义分位数：

- 总体分位数：对于一个分布函数 $F(x)$ ，则定义 q -分位数 (q -Quantiles) 为：

$$Q_q = F^{-1}(q), 0 < q < 1$$

- 回忆分布函数的定义为： $F(x) = P(X \leq x)$ ，因而 $q = F(Q_q) = P(X \leq Q_q)$

常用的分位数

由于 $q \in (0, 1)$ ，因而我们可以定义无数中分位数，常用的有：

- 中位数 (median) : $q = 0.5$ ，即有一半小于中位数，一半大于中位数
- 四分位数 (quartiles) : $q = 0.25/0.5/0.75$ ，三个四分位数将所有数据等分为四份
 - 四分位差：上四分位数减去下四分位数，通常用来度量离散程度，与方差类似
- 十分位数：9个分位数将数据分为十等分
- 百分位数 (percentiles) : 99个分位数将数据分为100等分

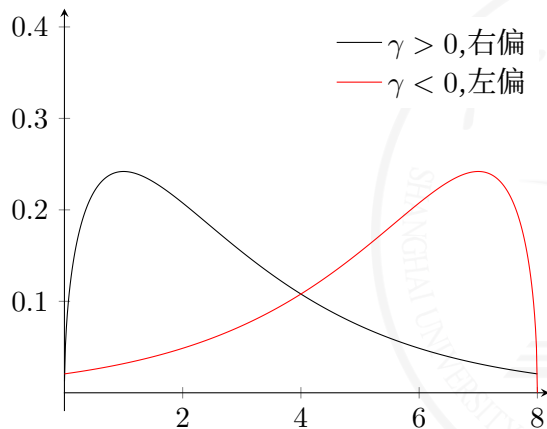
中位数的性质

中位数和样本均值都是数据集中趋势（或者平均水平的度量），但是具有和样本均值不同的性质：

- 均值容易受到异常值的影响，而中位数不容易受到异常值影响
 - 同样，四分位差相对于方差、标准差，也不容易受到异常值的影响
- 中位数、均值与偏态的关系：
 - 对于左偏分布，均值小于中位数
 - 对于右偏分布，均值大于中位数
 - 非参数偏度： $\gamma = \frac{\bar{x} - Q_{0.5}}{s}$
- 中位数对于单调变换保持不变，而均值对非线性单调变换不保证成立，即对于非线性的函数 $g(\cdot)$ ：

$$f[Q_q(X)] = Q_q(f(X)), \mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$$

偏态与中位数、均值



中位数、均值的非线性单调变换

中位数、均值的对数变换

如果有一组数据：1, 10, 100, 1000, 10000，那么均值为： $\bar{x} = 2222.2$ ，中位数为 $M_{\frac{1}{2}} = 100$ ，严重右偏

现在对这组数据进行对数变换，即 $\tilde{x}_i = \log_{10} x_i$ ，那么变换后的数据为：0, 1, 2, 3, 4，那么均值为： $\bar{\tilde{x}} = 2$ ，中位数为 $Q_{0.5}(\tilde{x}) = 2$

注意到，

$$\log_{10} \bar{x} \approx 3.347 > \overline{\log_{10} x_i} = 2$$

而：

$$\log_{10} Q_{0.5}(x) = 2 = Q_{0.5}(\log_{10} x)$$

中位数的计算

- 我们知道，期望是如下问题的解：

$$\mathbb{E}(x) = \arg \min_c \mathbb{E}(x - c)^2$$

而样本均值 $\bar{x}$ 相应的定义为：

$$\bar{x} = \arg \min_c \frac{1}{N} \sum_{i=1}^N (x_i - c)^2$$

- 而总体中位数 $Q_{0.5}$ 则是如下问题的解：

$$Q_{0.5} = \arg \min_c \mathbb{E}|x - c|$$

相应的，样本中位数可以通过解：

$$M = \arg \min_c \frac{1}{N} \sum_{i=1}^N |x_i - c|$$

得到。

中位数的计算

Why? 根据一阶条件：

$$\begin{aligned}
 0 &= \frac{\partial \mathbb{E} |X - c|}{\partial c} \\
 &= \frac{\partial \int_{\mathbb{R}} |x - c| dF(x)}{\partial c} \\
 &= \int_{\mathbb{R}} \frac{\partial |x - c|}{\partial c} dF(x) \\
 &= \int_c^{\infty} (-1) dF(x) + \int_{-\infty}^c 1 dF(x) \\
 &= -[1 - F(c)] + [F(c) - 0] \\
 &= -1 + 2F(c)
 \end{aligned}$$

从而 $c = F^{-1}(\frac{1}{2})$ ，即总体中位数。

q -分位数的计算

更进一步，对于任意的 q -分位数，可以通过解：

$$Q_q = \arg \min_c \mathbb{E} \psi_q(x - c)$$

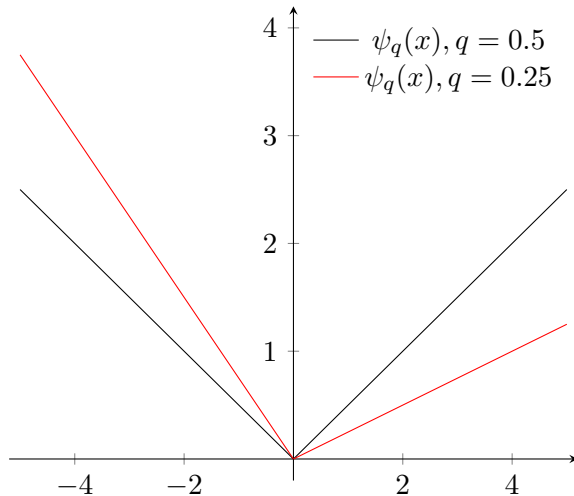
其中：

$$\psi_q(x) = x(q - \mathbb{1}\{x \leq 0\}) = \begin{cases} qx & x > 0 \\ (q-1)x & x \leq 0 \end{cases}$$

类似的，样本分位数可以通过解：

$$m_q = \arg \min_c \frac{1}{N} \sum_{i=1}^N \psi_q(x_i - c)$$

q -分位数的计算



条件分位数

- 现在，仿照条件期望函数的定义，我们可以定义总体的条件分位数函数（conditional quantile function），即

$$\mathbb{Q}_q(y_i|x_i) = \arg \min_h \mathbb{E} \psi_q(y_i - h(x_i))$$

- 与条件期望函数类似，条件分位数函数可以理解为给定 x_i 的情况下 y_i 的分位数，即

$$P(y_i \leq \mathbb{Q}_q(y_i|x_i) | x_i) = q$$

从而条件分位数函数描述了随着自变量 x_i 的变化， y_i 的条件分位数的变化。

分位数回归

- 如果假设 $\mathbb{Q}_q(y_i|x_i) = x_i'\beta_q$, 那么可以求解

$$\hat{\beta}_q = \arg \min_{\beta} \frac{1}{N} \sum_{i=1}^N \psi_q(y_i - x_i'\beta)$$

从而 $\widehat{\mathbb{Q}_q(y_i|x_i)} = x_i'\hat{\beta}_q$, 即为 (q -) 分位数回归 (quantile regression) .

- 特别的, 对于中位数, 可以通过解:

$$\hat{\beta}_{0.5} = \arg \min_{\beta} \frac{1}{N} \sum_{i=1}^N |y_i - x_i'\beta|$$

即为中位数回归 (median regression)

分位数回归

分位数回归的实现

在Stata中，可以使用qreg命令完成分位数回归的计算。比如在例消费的例子中，可以使用命令

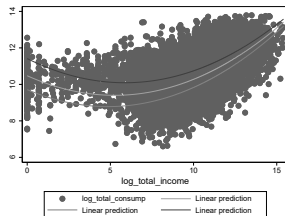
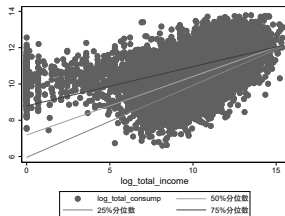
```
1 | qreg total_consump total_income, q(0.25)
```

进行25%的分位数回归。如果忽略q()选项，则默认进行中位数回归。

分位数回归

- 我们可以进行多次估计，在不同的分位数上进行分位数回归，从而可以观察 $y|x$ 的分布随着 x 的变化而变化的情况.
- 不过需要注意的是，原则上不同分位数的分位数回归函数是不能相交的，如果存在相交，则可能是条件分位数函数设定错误导致的。
 - `qreg_consump.do`

分位数回归



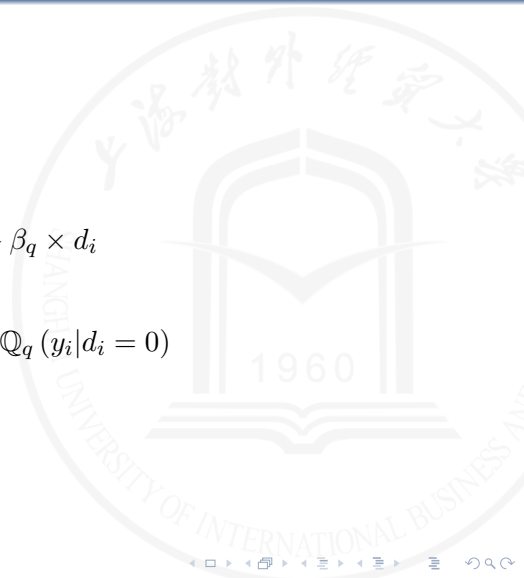
分位数回归：解释

- 分位数回归的解释：对于分位数回归：

$$Q_q(y_i|d_i) = \alpha + \beta_q \times d_i$$

那么

$$\beta_q = Q_q(y_i|d_i = 1) - Q_q(y_i|d_i = 0)$$



分位数回归：解释

性别收入差距

- 代码“qreg_with_dummy.do”使用分位数回归重新计算了性别收入差距. 可以计算数据中女性的中位数收入为30000元，男性的中位数收入为36000元，而中位数回归的结果 $\hat{\beta}_{0.5} = 6000$ 与以上结果完全等价，其他分位数同样可以验证.
- 如此， $\hat{\beta}_{0.5} = 6000$ 可以解释为男性的中位数收入比女性的中位数收入高6000元.
- 而 $\hat{\beta}_{0.75} = 9600$ 意味着男性的75%分位数收入比女性的75%分位数收入高9600元，以此类推.

分位数回归：随机系数的解释

或者，我们可以从随机系数（random coefficients）的角度来理解：

- 考虑如下数据生成过程：
 - $q_i \sim \mathcal{U}(0, 1)$
 - 如果给定 x_i , y_i 刚好处在 q_i 分位数上，而 q_i 分位数所对应的参数为 $\beta(q_i) = [\beta_k(q_i)]$ ，那么我们可以通过

$$y_i = \beta_1(q_i) + \beta_2(q_i) \cdot x_{i2} + \cdots + \beta_K(q_i) \cdot x_{iK} + e_i$$

- 来生成满足

$$\mathbb{Q}_q(y_i | d_i) = \beta_1(q_i) + \beta_2(q_i) \cdot x_{i2} + \cdots + \beta_K(q_i) \cdot x_{iK}$$

的数据。

- 需要保证 $\mathbb{Q}_q(y_i | d_i)$ 对 q 是单调的。

分位数回归：随机系数的解释

- qreg_simulate.do



分位数回归 v.s. 线性回归：异常值

qreg_outlier.do

