2017「应用微观计量经济学」暑期学校

计量经济学回顾

司继春

上海对外经贸大学

2017年8月

什么是计量经济学?

应用场景分类:

1. 微观计量
2. 宏观计量
3. 金融计量

工具方法分类:

1. 简约式（reduced form）
2. 结构式（structrual form）

微观计量经济学的任务

Heckman(2008)提出，因果推断的三个任务：

1. 定义问题、假说或者反事实——逻辑、理论与想象
2. 从总体数据中识别参数——如果没有抽样偏差，能否还原反事实？
3. 从实际数据中识别参数——统计理论，估计与检验

「Effect of Causes」 versus 「Causes of Effects」

基本工具：作为拟合的回归

如果有数据$(y_i, x_i')'$，令残差$e_i = y_i - x_i' b$，那么不同形式的回归：

1. 普通线性回归：

$$\min_b \sum_{i=1}^{N} e_i^2$$

2. 中位数回归：

$$\min_b \sum_{i=1}^{N} |e_i|$$

3. q-分位数回归：

$$\min_b \sum_{i=1}^{N} \psi_q (e_i)$$

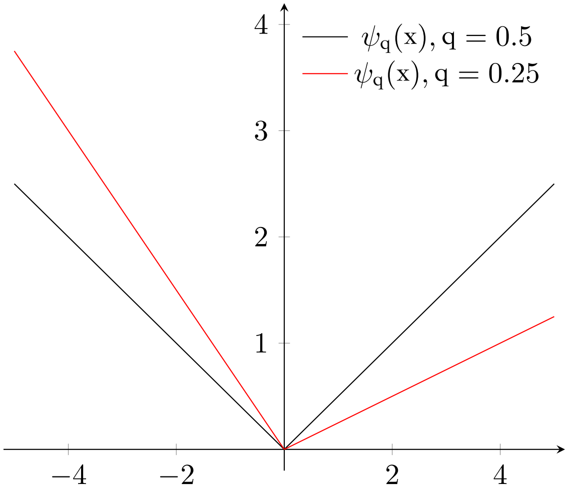
   其中

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

4. Probit、Logit回归

分位数回归

$\psi_q(x)$:



最小二乘法（OLS）

如果我们记：

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, X = [X_1, X_2, \cdots X_K] = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ X_{21} & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{NK} \end{bmatrix}$$

那么回归方程可以写作：

$$Y = Xb + e$$

而最小二乘法最小化：

$$\min_{b} (Y - Xb)' (Y - Xb)$$

从而得到：$\hat{b} = (X'X)^{-1} X'Y$。

最小二乘法的几何性质

根据OLS，由于$\hat{b} = (X'X)^{-1}X'Y$，因而Y的拟合值为：

$$\hat{Y} = X\hat{b} = X\left(X'X\right)^{-1}X'Y$$

记$P = X\left(X'X\right)^{-1}X'$，从而$\hat{Y} = PY$，即P做为一个线性变换，将向量Y变换到其拟合值$\hat{Y}$。
而残差为：

$$\hat{e} = Y - \hat{Y} = [I - P]\,Y$$

记$M = I - P$，从而$\hat{e} = MY$，即线性变换M将向量Y变换到其残差$\hat{e}$。

最小二乘法的几何性质

M和P矩阵具有以下性质：

$$P^2 = X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' = X\left(X'X\right)^{-1}X' = P$$

$$M^2 = (I - P)(I - P) = I - P - P + P^2 = I - P = M$$

如果矩阵A满足$A^2 = A$，那么我们称A为幂等矩阵（idempotent matrix）。可见P和M矩阵都是对称的幂等矩阵，或者投影矩阵。

最小二乘法的几何性质

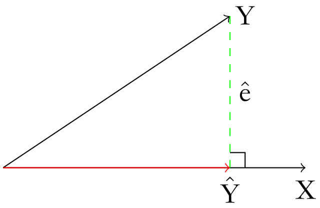
注意到：

$$\mathrm{PM} = \mathrm{P}(\mathrm{I} - \mathrm{P}) = \mathrm{P} - \mathrm{P}^2 = 0$$

因而如果计算Y的拟合值$\hat{\mathrm{Y}}$与残差$\hat{\mathrm{e}}$之间的内积，有：

$$\hat{\mathrm{Y}}'\hat{\mathrm{e}} = \mathrm{Y}'\mathrm{PMY} = 0$$

即拟合值$\hat{\mathrm{Y}}$与残差$\hat{\mathrm{e}}$正交。



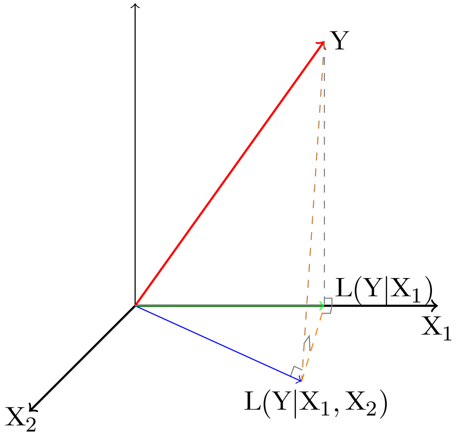
最小二乘法的几何性质

由于

$$\hat{Y} = PY = Xb = \sum_{j=1}^{k} b_j X_j$$

因而$\hat{Y}$是X_j的一个线性组合，且由于残差$\hat{e} = Y - \hat{Y}$与X平面正交，因而我们称$\hat{Y} = PY$为一个线性投影（linear projection）。

最小二乘法的几何性质



条件期望

以上讨论了向量空间中的投影。条件期望（conditional expectation）可以看成是无限维函数空间中的正交投影。
对于随机变量y和随机向量x，如果我们希望用x最优的预测y，可以设立如下目标函数：

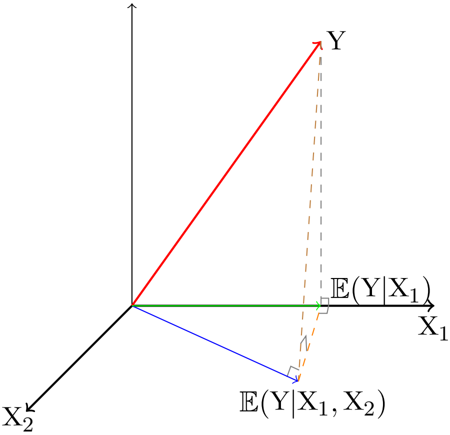
$$h^*(x) = \arg\min_{h \in \mathbb{H}} \mathbb{E}[y - h(x)]^2$$

其中h为任意的函数形式。
我们记以上最优函数$h^*(x)$为$\mathbb{E}(y|x)$，即随机变量y给定x的条件期望。

条件期望的性质

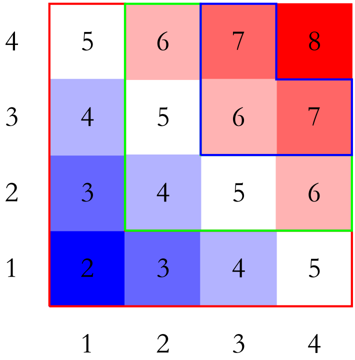
1. $\mathbb{E}\left(\mathrm{f}\left(\mathrm{x}\right)|\mathrm{x}\right) = \mathrm{f}\left(\mathrm{x}\right)$
2. $\mathbb{E}\left(\left[\mathrm{y} - \mathbb{E}\left(\mathrm{y}|\mathrm{x}\right)\right] \cdot \mathrm{g}\left(\mathrm{x}\right)\right) = 0$
3. $\mathbb{E}\left[\mathbb{E}\left(\mathrm{y}|\mathrm{x}_1, \mathrm{x}_2\right)|\mathrm{x}_1\right] = \mathbb{E}\left(\mathrm{y}|\mathrm{x}_1\right)$
4. $\mathbb{E}\left(\mathrm{f}\left(\mathrm{x}\right)\mathrm{y}|\mathrm{x}\right) = \mathrm{f}\left(\mathrm{x}\right)\mathbb{E}\left(\mathrm{y}|\mathrm{x}\right)$

迭代期望公式



条件期望

下图中代表两个四面骰子的结果，方框中代表两个骰子之和。
记x_1为第一个骰子的结果，x_2为第二个骰子的结果，
而y为两个骰子的和。



那么$\mathbb{E}\left(x_1|y=3\right)=?$

偏效应

在计量经济学中，由于维数问题（dimension problem）以及支撑问题（support problem），我们通常设定一个参数模型，将条件期望定义的函数范围限制在一个参数族内：

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^k} \left[y - h_\beta(x) \right]^2$$

其中$h_\beta(x)$为一个依赖于参数β的已知的函数形式。最常用的如线性回归：

$$h_\beta(x) = x'\beta$$

1. 与线性回归区别：总体 v.s. 样本
2. 如果令$y = x'\beta + u$，在这里$\mathbb{E}(u|x) = 0$是假设么？

偏效应

在参数模型下，偏效应为:

$$\frac{\partial \mathbb{E}\left(y|x\right)}{\partial x_j} = \frac{\partial h_\beta\left(x\right)}{\partial x_j}$$

1. 线性模型：$h_\beta\left(x\right) = x'\beta$，则偏效应为$\frac{\partial \mathbb{E}(y|x)}{\partial x_j} = \beta_j$
2. $h_\beta\left(x\right) = \beta_1 x_1 + \beta_2 x_1^2 + \widetilde{x}'\delta$，则$x_1$的偏效应为$\frac{\partial \mathbb{E}(y|x)}{\partial x_1} = \beta_1 + 2\beta_2 x_1$
3. $h_\beta\left(x\right) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \widetilde{x}'\delta$，则$x_1$的偏效应为$\frac{\partial \mathbb{E}(y|x)}{\partial x_1} = \beta_1 + \beta_3 x_2$
4. $h_\beta\left(x\right) = \exp\left\{\beta_1 x_1 + \beta_2 \ln x_2\right\}$，则$x_2$的偏效应为$\frac{\partial \mathbb{E}(y|x)}{\partial x_2} = \beta_2 \exp\left\{\beta_1 x_1 + \beta_2 \ln x_2\right\}\frac{1}{x_2}$。

偏效应

对于Probit/Logit模型：

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

则偏效应：

$$\frac{\partial\mathbb{E}(\mathrm{y}=1|\mathrm{x})}{\partial \mathrm{x}_j}=\frac{\partial \mathrm{F}(\mathrm{x}'\beta)}{\partial \mathrm{x}_j}=\mathrm{f}\left(\mathrm{x}'\beta\right)\beta_j$$

解释：当x增加一单位所带来的y的概率的变化，随着x的改变而改变。
平均偏效应：

$$\mathrm{APE}=\mathbb{E}\left[\mathrm{f}\left(\mathrm{x}'\beta\right)\beta_j\right]$$

而PEA（partial effects on average）为：

$$\mathrm{PEA}=\mathrm{f}\left[\mathbb{E}\left(\mathrm{x}'\beta\right)\right]\beta_j$$

几个独立性假定

1. $Y \amalg W \iff F_{W,Y}(w,y) = f_W(w) f_Y(y)$
2. $\mathbb{E}(Y|W) = \mathbb{E}(Y)$
3. $\mathrm{Cov}(Y,W) = 0$
4. $Y \amalg W|X \iff F_{W,Y|X}(w,y|x) = f_{W|X}(w|x) f_{Y|X}(y|x)$
5. $\mathbb{E}(Y|W,X) = \mathbb{E}(Y|X)$

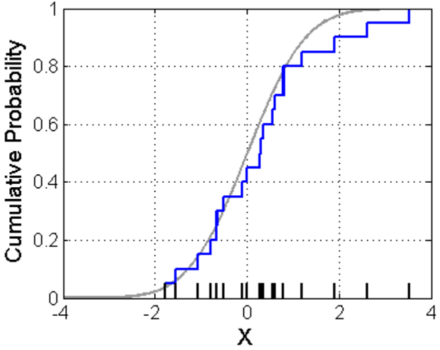
条件期望的估计

1. 非参数方法：核密度估计、Sieve估计
   1.1 优点：无函数形式假定
   1.2 缺点：维数问题、支撑问题
2. 参数方法：假定h (x)的函数形式
   2.1 优点：无维数问题、支撑问题
   2.2 缺点：需要假设函数形式

非参数核密度估计

经验分布函数:

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x_i \le x)$$



非参数核密度估计

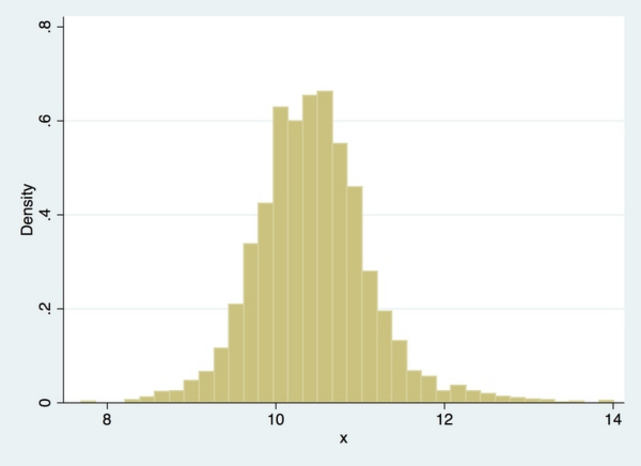
根据导数定义：

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$$

如果用经验分布函数替代分布函数，得到：

$$\begin{aligned}
\hat{f}(x) &= \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h} \\
&= \sum_{i=1}^{N} \frac{1(x-h \le x_i \le x+h)}{2Nh} \\
&= \frac{1}{Nh} \sum_{i=1}^{N} \frac{1}{2} \cdot 1\left(\frac{|x-x_i|}{h} \le 1\right)
\end{aligned}$$

非参数核密度估计



非参数核密度估计

然而以上估计不够光滑，注意到，如果记$K_0(t) = \frac{1}{2} \cdot 1(t < 1)$，那么上述估计可以写为:

$$\hat{f} = \frac{1}{Nh} \sum_{i=1}^{N} K_0\left(\frac{|x - x_i|}{h}\right)$$

可以替换$K_0(t)$为任意的密度函数，如正态分布密度函数，则:

$$\hat{f} = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{|x - x_i|}{h}\right)$$

仍然为密度函数。我们称$K(t)$为核函数。

1. 随着$N \to \infty$，$h \to 0$
2. h的选取：croww-validation/plug-in
3. bias-variance tradeoff

非参数回归

如果设定

$$y = g(x) + u$$

且$\mathbb{E}(u|x) = 0$，那么：

$$\mathbb{E}(y|x) = \int yf(y|x)\,dy = \frac{\int yf(y,x)\,dy}{f_X(x)}$$

如果使用对称的核函数估计其中的两个密度函数，则非参数回归可以写为：

$$\frac{\sum_i K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_i K\left(\frac{x-x_i}{h}\right)}$$

或者称核平滑（kernel smoothing）。

半参数回归

很多时候我们可以不完全非参数设定模型，而是使用：

$$y = x'\beta + g(z) + u$$

其中$\mathbb{E}(u|x,z) = 0$。根据迭代期望公式：

$$\mathbb{E}(y|z) = \mathbb{E}[\mathbb{E}(y|x,z)|z] = \mathbb{E}(x|z)'\beta + g(z)$$

以上两式相减：

$$y - \mathbb{E}(y|z) = [x - \mathbb{E}(x|z)]'\beta + u$$

1. 分别用y和x对z做非参数回归得到残差
2. 使用两个残差做线性回归

Simpson悖论

- 假设人群中有两类人，吃素的和吃肉的，记$d_i=1$为素食主义者，$d_i=0$为肉食者。记$x_i=1$为女性，$x_i=0$为男性。
- 假设是否吃素对寿命并没有影响，素食主义者的寿命为：

$$y_{1i}=70+10x_i+u_{1i}$$

而肉食者的寿命为：

$$y_{0i}=70+10x_i+u_{0i}$$

- 此时，由于$d_i \amalg (y_{0i},y_{1i})$，因而素食主义者与肉食主义者的寿命差别为0：

$$\mathbb{E}(y_{1i}|d_i=1)-\mathbb{E}(y_{0i}|d_i=0)=\mathbb{E}(y_{1i})-\mathbb{E}(y_{0i})=0$$

Simpson悖论

- 现在，假设素食主义者在男女性别中的比例是不一样的，假设女性中素食主义者的比例为75%，男性为20%，即：

$$
\begin{aligned}
\mathrm{P}\left(\mathrm{d_i}=1|\mathrm{x_i}=1\right)&=0.8\\
\mathrm{P}\left(\mathrm{d_i}=1|\mathrm{x_i}=0\right)&=0.1
\end{aligned}
$$

- 那么：

$$
\begin{aligned}
\mathrm{P}\left(\mathrm{x_i}=1|\mathrm{d_i}=1\right)&=\frac{\mathrm{P}\left(\mathrm{d_i}=1|\mathrm{x_i}=1\right)\mathrm{P}\left(\mathrm{x_i}=1\right)}{\mathrm{P}\left(\mathrm{d_i}=1|\mathrm{x_i}=1\right)\mathrm{P}\left(\mathrm{x_i}=1\right)+\mathrm{P}\left(\mathrm{d_i}=1|\mathrm{x_i}=0\right)\mathrm{P}\,(}\\
&=\frac{0.8}{0.8+0.1}=\frac{8}{9}
\end{aligned}
$$

同理：

$$
\mathrm{P}\left(\mathrm{x_i}=1|\mathrm{d_i}=0\right)=\frac{0.2}{0.2+0.9}=\frac{2}{11}
$$

Simpson悖论

- 此时，素食主义者的寿命：

$$
\begin{aligned}
\mathbb{E}\left(y_{1i}|d_i=1\right) &= \mathbb{E}\left(y_{1i}|d_i=1,x_i=1\right)P\left(x_i=1|d_i=1\right) \\
&+ \mathbb{E}\left(y_{0i}|d_i=1,x_i=0\right)P\left(x_i=0|d_i=1\right) \\
&= 80\times\frac{8}{9}+70\times\frac{1}{9}=78.9
\end{aligned}
$$

- 而同理，

$$
\begin{aligned}
\mathbb{E}\left(y_{0i}|d_i=0\right) &= \mathbb{E}\left(y_{0i}|d_i=0,x_i=1\right)P\left(x_i=1|d_i=0\right) \\
&+ \mathbb{E}\left(y_{0i}|d_i=0,x_i=0\right)P\left(x_i=0|d_i=0\right) \\
&= 80\times\frac{2}{11}+70\times\frac{9}{11}=71.8
\end{aligned}
$$

- 由于$d_i \amalg \left(y_{0i},y_{1i}\right)$不成立（Why?），因而看起来素食主义者更长寿。

Simpson悖论

- 然而此时，如果在相同性别组内进行比较，则可以得到素食无影响的结论：

$$\mathbb{E}\left(y_i|x_i=1,d_i=1\right)-\mathbb{E}\left(y_i|x_i=1,d_i=0\right)=0$$

$$\mathbb{E}\left(y_i|x_i=0,d_i=1\right)-\mathbb{E}\left(y_i|x_i=0,d_i=0\right)=0$$

- 在这里，尽管$d_i \amalg \left(y_{0i},y_{1i}\right)$不成立，但是$d_i \amalg \left(y_{0i},y_{1i}\right)|x_i$成立。

Simpson悖论

如果$d_i \amalg (y_{0i}, y_{1i}) | x_i$也不成立呢?

- 令$d_i = 1(x_i - u_{1i} + u_{0i} \geq 0)$。
- 在此设定下，即使给定x_i，d_i仍然与不可观测的u_{1i}, u_{0i}相关，而u_{1i}, u_{0i}则直接影响了y_{1i}, y_{0i}。
- 注意，如果$-u_{1i} + u_{0i}$比较高，则成为素食者的概率会提高，即如果一个人吃素食的寿命更高，反而选择素食的概率更低，那么:

$$
\begin{aligned}
&\mathbb{E}(y_i | x_i = 1, d_i = 1) - \mathbb{E}(y_i | x_i = 1, d_i = 0) \\
&\quad = \mathbb{E}(80 + u_{1i} | d_i = 1) - \mathbb{E}(80 + u_{0i} | d_i = 0) \\
&\quad = \mathbb{E}(u_{1i} | d_i = 1) - \mathbb{E}(u_{0i} | d_i = 0)
\end{aligned}
$$

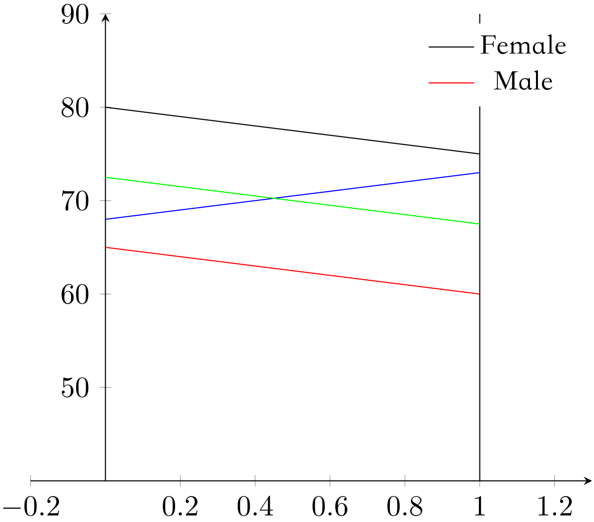
Simpson悖论

最终$\mathbb{E}(y_i|d_i=1)-\mathbb{E}(y_i|d_i=0)$符号取决于两种偏差：

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

1. 由于$d_i \amalg (y_{0i}, y_{1i})\,|x_i$不成立，因而$\mathbb{E}(y|x,d)\neq\mathbb{E}(y|x)$
2. $P(x=1|d=1), P(x=1|d=0)$，两个比例不相等。

Simpson悖论



线性回归

对于模型

$$y_i = x_i'\beta + u_i$$

如果假设:

$$\mathbb{E}\left(u_i | x_i\right) = 0$$

如何估计参数β?

线性回归

通常得到最小二乘的两个途径：

1. 最小二乘法
   - 强调拟合而并没有使用$\mathbb{E}(u_i|x_i) = 0$这一假设。
2. （拟）极大似然估计
   - 使用了比$\mathbb{E}(u_i|x_i) = 0$更强的假设，即$f(u_i|x_i)$的分布假设

第三个更常用的办法：矩估计。

线性回归

矩估计的步骤：

1. 根据模型得到一些总体矩的推论，并检查识别条件
   - $\mathbb{E}\left(\mathrm{u_i}|\mathrm{x_i}\right)=0 \Rightarrow \mathbb{E}\left(\mathrm{x_i u_i}\right)=\mathbb{E}\left[\mathbb{E}\left(\mathrm{x_i u_i}|\mathrm{x_i}\right)\right]=\mathbb{E}\left[\mathrm{x_i}\mathbb{E}\left(\mathrm{u_i}|\mathrm{x_i}\right)\right]=0$。
   - 从而：$\mathbb{E}\left[\mathrm{x_i}\left(\mathrm{y_i}-\mathrm{x_i'}\beta\right)\right]=0$
2. 将总体矩用样本矩进行替代（$\sum$替代$\mathbb{E}$），并将参数写成待估计形式
   - $\sum_{\mathrm{i}}\left[\mathrm{x_i}\left(\mathrm{y_i}-\mathrm{x_i'}\hat{\beta}\right)\right]=0$
3. 解以上方程（组）得到估计
   - $\sum_{\mathrm{i}}\left[\mathrm{x_i}\left(\mathrm{y_i}-\mathrm{x_i'}\hat{\beta}\right)\right]=\sum_{\mathrm{i}}\mathrm{x_i y_i}-\sum_{\mathrm{i}}\mathrm{x_i x_i'}\hat{\beta}=0$
   - $\hat{\beta}=\left(\sum_{\mathrm{i}}\mathrm{x_i x_i'}\right)^{-1}\sum\mathrm{x_i y_i}$

线性回归

回归的本质：比较

- 如果令$x_i = 0/1$，那么我们的矩条件为：

$$\begin{aligned} \mathbb{E}\left[x_i\left(y_i - \alpha - \beta x_i\right)\right] &= 0 \\ \mathbb{E}\left[y_i - \alpha - \beta x_i\right] &= 0 \end{aligned}$$

  解得：

$$\beta = \mathbb{E}\left(y_i | x_i = 1\right) - \mathbb{E}\left(y_i | x_i = 0\right)$$

- 使用样本矩替代，即：

$$\hat{\beta} = \bar{y}_1 - \bar{y}_0$$

线性回归

分步回归：variation究竟来自哪里？

- 对于线性回归模型：$y_i = x_i'\beta + z_i'\delta + u_i$
- β的OLS估计与以下步骤得到的回归系数是等价的：
  1. 分别用x和y对z做回归得到残差e_x, e_y
  2. 用e_y对e_x做回归，得到β的估计。

启示：

1. 只有z对x和y不能解释的那部分才是可用的variation
2. 为了得到β的一致估计，只需要$u \perp x|z$。

线性回归

如果$y_i = x_i'\beta + z_i'\delta + u_i$，其中$z_i = 0/1$。按照分步回归的步骤:

1. 用x和y对z做回归，得到残差。注意由于z为0-1变量，因而实际结果是x和y被分组去均值。
2. 用e_y对e_x做回归，得到β的估计。

注意在以上步骤中，实际使用的variation是被去掉组均值的残差，因而不同组别之间没有variation，是一个「组内」估计量。

线性回归

异方差的处理:

- 从参数估计的角度讲，没有必要对异方差做特殊处理，OLS的估计是一致的，尽管不再是有效的。
- 从假设检验的角度讲，不能忽略异方差的影响，异方差导致参数估计的s.e.不正确

解决方法：White异方差:

$$\mathrm{Var}\left(\hat{\beta}\right) = \left[\mathrm{X}'\mathrm{X}\right]^{-1}\left[\sum_{i=1}^{N} \hat{\mathrm{u}}^2 \mathrm{x}_i \mathrm{x}_i'\right]\left[\mathrm{X}'\mathrm{X}\right]^{-1}$$

线性回归

很多时候我们使用的是分组数据:

$$\mathrm{y_{ig}} = \mathrm{x}'_{\mathrm{ig}}\beta + \mathrm{u_{ig}}$$

在组内，很多时候$\mathrm{u_{ig}}$之间是相关的:

$$\mathbb{E}\left(\mathrm{u_{ig}u_{jg'}}\right) = \begin{cases} \sigma_{\mathrm{(ij)g}} & \mathrm{g} = \mathrm{g}' \\ 0 & \mathrm{g} \neq \mathrm{g}' \end{cases}$$

此时，仅仅做异方差文件的s.e.是不够的。

线性回归

可能的原因：组内某些不能观察到的变量对y_{ig}有影响。

- 例如，如果y_{ig}是班级内学生的成绩，那么g是班级，教师质量等都可能影响成绩。

很多时候不能单纯使用group dummy来解决

- 例如，学生努力程度的peer effect

线性回归

使用：

$$\mathrm{Var}\left(\hat{\beta}\right) = [\mathrm{X}'\mathrm{X}]^{-1}\left[\sum_{\mathrm{g}} \mathrm{x}_{\mathrm{g}}'\hat{\mathrm{u}}_{\mathrm{g}}\hat{\mathrm{u}}_{\mathrm{g}}'\mathrm{x}_{\mathrm{g}}\right][\mathrm{X}'\mathrm{X}]^{-1}$$

$$\begin{bmatrix}
\sigma^2_{(11)1} & \cdots & \sigma_{(1N_1)1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & & \vdots \\
\sigma^2_{(N_1 1)1} & & \sigma_{(N_1 N_1)1} & 0 & & 0 & & 0 & & 0 \\
0 & \cdots & 0 & \sigma^2_{(11)2} & \cdots & \sigma_{(1N_2)2} & & \vdots & & \vdots \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & & & \\
0 & \cdots & 0 & \sigma_{(N_2 1)2} & \cdots & \sigma^2_{(N_2 N_2)2} & & & & \\
 & & & & & & \ddots & \vdots & & \vdots \\
 & & & & & & & \sigma^2_{(11)g} & \cdots & \sigma_{(1N_g)g} \\
 & & & & & & & \vdots & \ddots & \vdots \\
 & & & & & & & \sigma_{(N_g 1)g} & \cdots & \sigma^2_{(N_g N_g)g}
\end{bmatrix}$$

线性回归

不止一组的情况:

- 如果是嵌套的，比如学校、班级，应该cluster在更高的一级
- 如果不是嵌套的:
  - 控制一组的固定效应，cluster另外一组
  - two-way cluster，Cameron, Gelbach and Miller, 2006
- 现实中:
  - reg中的cluster()只能cluster一组
  - reghdfe

线性回归

一般来说，cluster的标准误更大，因而参数的置信区间更大，越容易不显著。cluster标准误比i.i.d的标准误扩大了：

$$\sqrt{1+\rho_{\mathrm{x}}\rho_{\mathrm{u}}\left(\bar{\mathrm{N}}-1\right)}$$

其中ρ_{x}为解释变量的组内相关性，ρ_{u}为误差项组内相关性，而$\bar{\mathrm{N}}$为平均的组大小。

内生性问题

在线性回归中，我们假设了$\mathbb{E}(\mathrm{u}|\mathrm{x})=0$，从而得到了最小二乘估计。当这一假定被违背时，我们称为有内生性问题。

1. 微观、宏观中的「内生变量」与计量中的内生性的联系区别?
2. 内生性可能的原因:
   - 互为因果
   - 遗漏变量
   - 度量误差
   - 自选择
   - ……

工具变量

解决方案：找到外生的影响x但是不会影响y的变量，即工具变量z。

如果我们要估计的方程为：

$$\mathrm{y} = \alpha + \beta \mathrm{y}_1 + \mathrm{u}$$

然而$\mathrm{Cov}\,(\mathrm{x}, \mathrm{u}) \neq 0$，但是我们可以找到一个z，使得$\mathrm{Cov}\,(\mathrm{z}, \mathrm{u}) = 0$，且：

$$\mathrm{y}_1 = \gamma + \delta \mathrm{z} + \mathrm{e}$$

我们把以上方程带入结构式，得到：

$$\begin{aligned} \mathrm{y} &= \alpha + \beta\left(\gamma + \delta \mathrm{z} + \mathrm{e}\right) + \mathrm{u} \\ &= \left(\alpha + \beta\gamma\right) + \beta\delta \mathrm{z} + \mathrm{u} + \beta \mathrm{e} \end{aligned}$$

我们称上式为简约式。

工具变量

现在两个式子:

$$x = \gamma + \delta z + e$$

以及

$$\begin{aligned} y &= (\alpha + \beta\gamma) + \beta\delta z + u + \beta e \\ &= \gamma^* + \delta^* z + v \end{aligned}$$

都不存在内生性问题，因而我们可以一致估计$\gamma, \delta, \gamma^*, \delta^*$。进而，可以得到$\beta$:

$$\beta = \frac{\delta^*}{\delta} = \frac{\beta\delta}{\delta}$$

假设:

1. z与u不相关
2. 分母（δ）不为0。

工具变量

一般情况下，如果估计方程：

$$y = \alpha y_1 + x'\beta + u$$

其中y_1为关心的内生变量，即$\mathrm{Cov}(y_1, u) \neq 0$，但是存在工具变量$z$，使得$\mathbb{E}(u|z) = 0$，且$\mathbb{E}(y_1|x, z) \neq \mathbb{E}(y_1|x)$，那么$y_1$的简约式可以写为：

$$y_1 = z'\gamma + x'\delta + v$$

那么可以通过工具变量达到对α的识别。

工具变量

工具变量的理解：以联立方程为例
如果存在互为因果的两个内生变量y_1, y_2，满足：

$$y_1 = \alpha_1 y_2 + x_1\beta_1 + u_1$$

$$y_2 = \alpha_2 y_1 + x_2\beta_2 + u_2$$

联立两个方程，可以得到：

$$\begin{aligned}
y_2 &= \alpha_2\left(\alpha_1 y_2 + x_1\beta_1 + u_1\right) + x_2\beta_2 + u_2 \\
&= \alpha_1\alpha_2 y_2 + \alpha_2 x_1\beta_1 + x_2\beta_2 + \alpha_2 u_1 + u_2 \\
&= \frac{\alpha_2}{1-\alpha_1\alpha_2}x_1\beta_1 + \frac{1}{1-\alpha_1\alpha_2}x_2\beta_2 + \frac{\alpha_2 u_1 + u_2}{1-\alpha_1\alpha_2} \\
&\triangleq x'\delta + v
\end{aligned}$$

使用决定y_2、但同时不决定y_1的外生的扰动，即包含在x_2而不包含在x_1的变量(z)进行识别。

工具变量

估计：两阶段最小二乘法（2SLS）

1. 第一阶段回归，对y_2的简约式进行回归：

$$y_1 = z'\gamma + x'\delta + v$$
$$\hat{y}_1 = z'\hat{\gamma} + x'\hat{\delta}$$

   或者$\hat{Y}_1 = P_Z Y_1$，其中$Z = (Z, X)$，$P_Z = Z\left(Z'Z\right)^{-1} Z'$。

2. 第二阶段回归，用$\hat{y}_1$代替y_1：

$$y = \hat{\alpha}\hat{y}_1 + x'\hat{\beta} + \hat{e}$$

或者：$\left(\hat{\alpha}', \hat{\beta}'\right)' = (XP_Z Z)^{-1}(XP_Z Y)$，其中$X = (Y_1, X)$。
注意：不要手动算两阶段最小二乘，差别：在算标准误的时候用$\hat{e}$还是$\hat{u} = y - \hat{\alpha}y_1 + x'\hat{\beta}$?

工具变量

工具变量的理解：我们使用了那些variation？

1. 第一阶段回归：
$$\hat{y}_1 = z'\hat{\gamma} + x'\hat{\delta}$$
2. 第二阶段回归：
$$y = \hat{\alpha}\hat{y}_1 + x'\hat{\beta} + \hat{e}$$

考虑使用分步回归，即先用$\hat{y}_1$对x做回归得到残差，等价于用$z'\hat{\gamma}$对x做回归，得到残差，再用y对以上残差做回归。
只使用了单纯因为工具变量而导致的y_1的变化的部分variation。
因而需要考虑清楚工具变量如何影响内生变量。
例：用电供给，用天气做IV或者用大宗商品价格做IV？

GMM

广义矩估计（Generalized Method of Moments）提供了一个简单实用的估计、检验框架。

- 依赖于矩条件（Moment Condition）
- 几乎所有的经典计量方法都可以看成是GMM的特例，如OLS、2SLS等
- 特别的，绝大多数情况下，MLE也可以看成是GMM的特例

GMM

考虑一个简单的求正态分布参数的问题。如果$\mathrm{x_i} \sim \mathrm{N}\left(\mu, \sigma^2\right)$，为了估计$\mu, \sigma^2$，我们可以：

1. MLE：最大化

$$\sum_{\mathrm{i}} \left[-\frac{1}{2} \ln\left(2\pi\right) - \ln\sigma - \frac{\left(\mathrm{x_i} - \mu\right)^2}{2\sigma^2} \right]$$

2. 矩估计（Method of Moments）：由于$\mathbb{E}\left(\mathrm{x_i}\right) = \mu$，$\mathbb{E}\left(\mathrm{x_i^2}\right) = \sigma^2 + \mu^2$，那么解方程：

$$\begin{cases} \frac{1}{\mathrm{N}} \sum_{\mathrm{i}} \left(\mathrm{x_i} - \hat{\mu}\right) & = 0 \\ \frac{1}{\mathrm{N}} \sum_{\mathrm{i}} \left(\mathrm{x_i^2} - \hat{\mu}^2 - \hat{\sigma}^2\right) & = 0 \end{cases}$$

GMM

然而我们知道，对于正态分布而言，不仅仅存在一阶矩、二阶矩，还有三阶矩、四阶矩以及更高阶矩，比如$\mathbb{E}\left(\mathrm{x}_{\mathrm{i}}^{3}\right)=\mu^{3}+3\mu\sigma^{2}$，如果联立：

$$\begin{cases}\frac{1}{\mathrm{N}}\sum_{\mathrm{i}}\left(\mathrm{x}_{\mathrm{i}}-\hat{\mu}\right) & =0\\ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}}\left(\mathrm{x}_{\mathrm{i}}^{2}-\hat{\mu}^{2}-\hat{\sigma}^{2}\right) & =0\\ \frac{1}{\mathrm{N}}\sum_{\mathrm{i}}\left(\mathrm{x}_{\mathrm{i}}^{2}-\hat{\mu}^{3}-3\hat{\mu}\hat{\sigma}^{2}\right) & =0\end{cases}$$

两个未知数、三个方程，无解。

GMM

其他矩估计：

1. 线性回归，矩条件：$\mathbb{E}(x_i u_i) = 0$，从而矩估计为：

$$\frac{1}{N}\sum_i x_i u_i = \frac{1}{N}\sum_i x_i \left(y_i - x_i'\hat{\beta}\right) = 0$$

   由于x_i为k维向量，待估参数β也是k维向量，从而可以解出估计参数。

2. 工具变量，如果记$z^* = (z', x')'$，即所有外生变量，那么矩条件：$\mathbb{E}(z_i^* u_i) = 0$，矩估计为：

$$\frac{1}{N}\sum_i z_i^* \left(y_i - \alpha y_{1i} - x_i'\hat{\beta}\right) = 0$$

   如果工具变量（包含x）个数L等于待估系数个数（x的维数加上内生变量个数）K，那么仍然有解。

3. 如果$L > K$，无解。

前两种情况：恰好识别；第三种情况：过度识别（overidentification）

GMM

在一定范围内，工具变量（矩条件）越多，提供的variation也越多，所以我们希望尽可能的用比较多的工具变量（矩条件）。如果我们有总体矩条件

$$\mathbb{E}\left[\mathrm{m}\left(\mathrm{w_i}, \theta\right)\right] = 0$$

其中$\mathrm{m}\left(\cdot\right)$为$\mathrm{L} \times 1$维，即有L个矩条件，$\theta$为K维，即有K个待估计参数。

1. 想法1：最小化

$$\min_{\theta} \sum_{\mathrm{i}} \mathrm{m}\left(\mathrm{w_i}, \theta\right)' \mathrm{m}\left(\mathrm{w_i}, \theta\right) = \sum_{\mathrm{i}} \left[\sum_{\mathrm{l}=1}^{\mathrm{L}} \mathrm{m_l}\left(\mathrm{w_i}, \theta\right)^2\right]$$

2. 想法2：继续推广，最小化：

$$\min_{\theta} \left[\sum_{\mathrm{i}} \mathrm{m}\left(\mathrm{w_i}, \theta\right)\right]' \mathrm{W} \left[\sum_{\mathrm{i}} \mathrm{m}\left(\mathrm{w_i}, \theta\right)\right]$$

其中W为$\mathrm{L} \times \mathrm{L}$的正定的权重矩阵。

GMM

- 给定任意的θ，目标函数总是≥ 0。
- 给定任意的正定矩阵W，当且仅当每个矩$\sum_{\mathrm{i}} \mathrm{m_i}\left(\mathrm{w_i}, \theta\right)$都等于0时，目标函数等于0。
- 因而以上目标函数保证每个矩条件都充分贴近与0，
- 识别条件：存在唯一解使得

$$\mathbb{E}\left[\mathrm{m}\left(\mathrm{w_i}, \theta\right)\right] = 0$$

GMM

如果给定任意一个权重矩阵W，都可以得到不同的估计结果：

- 理论上，给定任意的正定权重矩阵W，都可以得到一致的结果
- 然而任意的正定权重矩阵所得到的估计结果渐进方差不相同
- 存在一个最优权重矩阵W^*使得估计结果方差最小：

$$\mathrm{W}^* = \left(\mathbb{E}\left[\mathrm{m}\left(\mathrm{w_i}, \theta\right)\mathrm{m}\left(\mathrm{w_i}, \theta\right)'\right]\right)^{-1}$$

- 可以使用：

$$\hat{\mathrm{W}} = \left(\sum_{\mathrm{i}=1}^{\mathrm{N}}\left[\mathrm{m}\left(\mathrm{w_i}, \theta\right)\mathrm{m}\left(\mathrm{w_i}, \theta\right)'\right]\right)^{-1}$$

GMM

实际操作中：

1. 使用任意的正定矩阵（如单位阵）作为权重矩阵进行估计，得到第一次估计$\hat{\theta}_0$
2. 将$\hat{\theta}_0$带入

$$\hat{W}_1 = \left(\sum_{i=1}^{N} \left[m\left(w_i, \theta\right) m\left(w_i, \theta\right)' \right] \right)^{-1}$$

   得到权重矩阵，再次进行估计（two-step）
3. 或者，重复以上步骤直至收敛。

- Stata: gmm命令；Matlab：GMM工具箱

GMM

过度识别检验：

- 在$L > K$的情形下，我们可以检验矩条件（工具变量）是否有效。
- 可以证明，当使用最优权重矩阵时，GMM目标函数渐进服从χ^2分布，自由度为$L - K$：

$$\left[\sum_i m\left(w_i, \hat{\theta}\right)\right]' \left(\sum_{i=1}^{N}\left[m\left(w_i, \hat{\theta}\right) m\left(w_i, \hat{\theta}\right)'\right]\right)^{-1} \left[\sum_i m\left(w_i, \hat{\theta}\right)\right] \overset{a}{\sim} \chi^2\left(L - K\right)$$

- Hansen's J-statistics

GMM

例子:

1. 工具变量，矩条件为：$\mathbb{E}\left(z_i u_i\right)=0$，则GMM目标函数：

$$\min \left[\sum_i z_i\left(y_i-\tilde{x}_i^{\prime} \tilde{\beta}\right)\right]^{\prime} W\left[\sum_i z_i\left(y_i-\tilde{x}_i^{\prime} \tilde{\beta}\right)\right]$$

   如果取$W=Z^{\prime} Z$，则得到了两阶段最小二乘（2SLS）。
2. 在同方差假定下，以上的权重矩阵为最优权重矩阵。
3. 当同方差假定不满足时，以上的权重矩阵不是最优权重矩阵。

GMM

例子：动态面板模型：

$$y_{it} = \alpha y_{i,t-1} + x_{it}'\beta + \eta_i + u_{it}$$

为了消除固定效应：

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \Delta x_{it}'\beta + \Delta u_{it}$$

然而，$\mathbb{E}\left(\Delta y_{i,t-1} \cdot \Delta u_{it}\right) = \mathbb{E}\left(y_{i,t-1} - y_{i,t-2}\right)\left(u_{it} - u_{i,t-1}\right) \neq 0$
工具变量：$y_{i,t-3}, y_{i,t-4}, \ldots, x_{i,t-1}, \cdots$

工具变量

控制函数法：如果估计方程：

$$y = \alpha y_1 + x'\beta + u$$

以及y_1的简约式：

$$y_1 = z'\gamma + x'\delta + v$$

在简约式中，y_1被分解为两部分：

1. $z'\gamma + x'\delta$是与u不相关的部分
2. 如果y_1与u相关，则相关性必在v里面。

工具变量

因而，如果假定：

$$u = \rho v + e$$

那么待估计方程变为：

$$y = \alpha y_1 + x'\beta + \rho v + e$$

因而如果观察到v，就可以直接控制「内生性」v从而达到识别。

工具变量

估计步骤：

1. 估计简约式$y_1 = z'\gamma + x'\delta + v$得到$\hat{v}$
2. 将$\hat{v}$带入主方程，使用回归：

$$y = \alpha y_1 + x'\beta + \rho\hat{v} + e$$

进行估计。

检验内生性：$H_0 : \rho = 0$。

工具变量

- 在线性模型下，控制函数法与2SLS等价
- 控制函数法在很多非线性模型下仍然可以使用，而2SLS不可以。
  - Probit模型：

$$d_i = 1\left(\alpha y_{1i} + x_i'\beta + u_i \geq 0\right)$$

$$y_{1i} = z'\gamma + x'\delta + v$$

  - 首先估计简约式方程，得到$\hat{v}$，进而回归：

$$d_i = 1\left(\alpha y_{1i} + x_i'\beta + \hat{v}_i + e_i \geq 0\right)$$

工具变量

- 或者，可以使用极大似然估计，即假设(e_i, v_i)的联合（正态）分布，进而：

$$f(y_i, y_{1i}|x_i, z_i) = f(y_i|y_{1i}, x_i, z_i, v_i) \cdot f(v_i|x_i, z_i)$$

如果假定(e_i, v_i)为联合正态分布，即得到有限信息极大似然估计（LIML）。

- 其他估计方法：k–class估计量

工具变量

工具变量的两个假定：

1. 与误差项不相关——Hansen's J test
2. 与内生变量高度相关

如果第二项假定不满足？

工具变量

以上得知，对于要估计的方程：

$$y = \alpha + \beta y_1 + u$$

以及第一阶段：

$$y_1 = \gamma + \delta z + e$$

估计可以写为：

$$\hat{\beta}_{2SLS} = \frac{\mathrm{Cov}(y, z)}{\mathrm{Cov}(y_1, z)}$$

如果分母趋向于0，IV估计的结果会非常不稳定——弱工具。

工具变量

- Angrist and Krueger (1991)研究教育回报时，使用了出生季节与州、年份的乘积作为上学时间的工具变量
- Bound, Jaeger and Baker (1996)对此研究做出了批评，发现随机生成一组出生季节仍然能得到相似的结论。
- 工具变量太多导致估计结果偏向OLS的估计结果

工具变量

- 对于少量弱工具：
  - 诊断弱工具
  - 使用LIML，并调整置信区间。
- 对于Many instruments：
  - 2SLS表现非常差
  - 使用LIML，或者REQML(Chamberlain and Imbense, 2004)

工具变量

弱工具诊断：

- 第一阶段F – value，一般在一个内生变量一个工具变量的情况下，大于10
- Cragg-Donald 统计量（F – value推广，当只有一个内生变量时就是F – value）
- Stock and Yogo (2005)提出了针对Cragg-Donald 统计量的临界值

置信区间调整：

- Anderson and Rubin (1949)

工具变量

```
Underidentification test (Anderson canon. corr. LM statistic):           0.124
                                         Chi-sq(1) P-val =    0.7249
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                0.122
Stock-Yogo weak ID test critical values: 10% maximal IV size            16.38
                                         15% maximal IV size             8.96
                                         20% maximal IV size             6.66
                                         25% maximal IV size             5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.000
                                         (equation exactly identified)
------------------------------------------------------------------------------
```