Bartik工具变量与模拟的工具变量

慧航

2025年11月

Bartik工具

Bartik工具变量，也称为移动-份额工具（shift-share instruments）最初由Bartik（1991）使用并讨论，在Blanchard和Katz（1992）之后更加流行，目前已近被应用于很多领域：

- 移民：Altonji和Card（1991）、Card（2001）
- 银行借贷：Amiti和Weinstein（2018）、Greenstone，Mas和Nguyen（2015）
- 市场规模与创新：Acemoglu和Linn（2004）
- 贸易：Autor, Dorn和Hanson（2013,2018）、Piveteau和Smagghue（2017）、de Roux等（2017）
- 国外援助：Nunn和Qian（2014）
- 自动化：Acemoglu和Restrepo（2017）

基本设定

基本（简单）设定

- 结构方程：

$$y_i = \beta_0 + \beta w_i + u_i$$

其中$i = 1, ..., N$为个体（通常为地区），
  - y_i为地区的被解释变量，为了保证接下来的外生性，通常被解释变量为相对值（比如差分、增长率等）
  - $\mathbb{C}(w_i, u_i) \neq 0$从而有内生性
- 假设w_i有如下的结构：

$$w_i = \sum_{m=1}^{M} z_{im} g_{im}$$

其中z_{im}为i地区第m个成分的份额，g_{im}为增长率

基本设定

估计劳动供给弹性

- y：工资增长率
- w：就业增长率
- z_{im}：i地区行业m的份额
- g_{im}：i地区行业m的工资增长率

估计移民的影响

- y：本地和移民工资差异
- w：移民增长率
- z_{im}：i地区来自国家m的移民份额
- g_{im}：i地区自国家m的人口增长率

Bartik工具的构建

- 现在假定：

$$g_{im} = g_m + \tilde{g}_{im}$$

其中g_m为第m个成分（行业、国家等）的平均增长率

- 构建Bartik工具变量：

$$B_i = \sum_{m=1}^{M} z_{im} g_m$$

- 问题：何时B_i是一个合格的工具变量？两个条件：
  - $\mathbb{C}\left(B_i, w_i\right) \neq 0$
  - $\mathbb{C}\left(B_i, u_i\right) = 0$
- 两只文献：
  - Goldsmith-Pinkham，Sorkin和Swift（2020）研究了份额z_{im}的外生性
  - Borusyak、Hull和Jaravel（2022）则研究了冲击g_m的外生性

一个特例：$M=2$

- 如果$M=2$，那么：

$$B_i = z_{i1}g_1 + z_{i2}g_2$$

由于$z_{i1}+z_{i2}=1$，从而：

$$B_i = g_2 + (g_1 - g_2)\, z_{i1}$$

- 第一阶段：

$$\begin{aligned} w_i &= \gamma_0 + \gamma B_i + \epsilon_i \\ &= \underbrace{\gamma_0 + \gamma g_2}_{\text{常数项}} + \underbrace{\gamma\,(g_1 - g_2)}_{\text{系数}} z_{i1} + \epsilon_i \end{aligned}$$

- 相关性要求$g_1 - g_2 \neq 0$：$g_1 - g_2$看做是某一个“政策”变动，

一个特例：$M = 2$

- 真正的工具：z_{i1}！
  - z_{i1}可以看做是政策的暴露水平
  - 问题：z_{i1}会不会直接影响y_i?
- 一般而言，无法排除z_{i1}作为份额与y_i的直接关系
  - 在劳动供给弹性的例子中，z_{i1}为某个行业的就业份额，y_i为收入，显然这是长期均衡共同决定的
  - 在移民的例子中，不同国家的收入不一样，显然当地收入也会影响不同国家的移民比例
  - 事情不一定这么悲观：y_i并不是收入，而是收入增长率（变动，而非水平），外生性也许会满足

M不受限、截面

- 现在考虑控制变量：

$$y_i = \beta_0 + \beta w_i + x_i' \eta + u_i$$

记$Y^{\perp} = Y - \mathbb{L}(Y|X)$，$W^{\perp} = W - \mathbb{L}(W|X)$

- 如果有M个行业，那么Bartik工具的2SLS估计：

$$\hat{\beta}^{\text{Bartik}} = \frac{B'Y^{\perp}}{B'W^{\perp}}$$

- 同时，如果使用$M-1$个份额作为工具变量，那么：

$$\hat{\beta}^{\text{GMM}} = \frac{W^{\perp\prime} Z \Xi Z' Y^{\perp}}{W^{\perp\prime} Z \Xi Z' W^{\perp}}$$

其中Ξ为权重矩阵

M不受限、截面

定理

定理：如果$\Xi = GG'$，那么$\hat{\beta}^{Bartik} = \hat{\beta}^{GMM}$

启示

Bartik工具变量估计的结果等价于使用份额z_{im}作为工具变量，并没有使用g_m的外生性！

面板情形

- 对于面板设定（首先忽略控制变量）：

$$y_{it} = \beta_t + \beta w_{it} + u_{it}$$

其中β_t为时间固定效应

- Bartik工具：

$$B_{it} = \sum_{m=1}^{M} z_{im0} g_{mt}$$

其中z_{im0}为初始的份额，而增长率：

$$g_{imt} = g_{mt} + \tilde{g}_{imt}$$

特例：$M=2, T=2$

- 此时工具：

$$B_{it}=g_{1t}z_{i10}+g_{2t}z_{i20}=g_{2t}+\left(g_{1t}-g_{2t}\right)z_{i10}$$

- 第一阶段：

$$\begin{aligned} w_{it} &= \gamma_t+\gamma B_{it}+\epsilon_{it} \\ &= \underbrace{\gamma_t+\gamma g_{2t}}_{\tilde{\tau}_t}+\gamma\left(g_{1t}-g_{2t}\right)z_{i10}+\epsilon_{it} \end{aligned}$$

如果将其用两期的虚拟变量写出来：

$$w_{it}=\tilde{\tau}_t+\underbrace{\gamma\left(g_{11}-g_{21}\right)}_{\tilde{\gamma}_1}\mathbb{1}\left\{t=1\right\}z_{i10}+\underbrace{\gamma\left(g_{12}-g_{22}\right)}_{\tilde{\gamma}_2}\mathbb{1}\left\{t=2\right\}z_{i10}+\epsilon_{it}$$

其中$\tilde{\gamma}_1,\tilde{\gamma}_2$为份额$z_{i10}$与时间固定效应的交互项的系数

- 等价关系：特定权重矩阵下，Bartik工具的估计结果与使用“份额与时间固定效应的交互”作为工具变量是等价的。

Bartik与DID

- 如果将$g_{1t}-g_{2t}$视为政策冲击，而z_{i10}视为政策暴露水平，以上设定与DID关系密切！
  - 具有高（$m=1$）份额的地区与低（$m=1$）份额的地区相比，对于w的增长率冲击$g_{1t}-g_{2t}$的反应是不是不同的?
  - 如果高低地区的w反应不同、y反应也不同，那么推断$w->y$有影响
- 问题：平行趋势？分组变量（z_{i10}）外生吗?
  - 如果存在政策冲击，比如$g_{11}-g_{21}=0$，那么可以检验z_{i10}是否对事前的y有影响：平行趋势检验
  - 直觉：如果$g_{11}-g_{21}=0$，且z_{i10}不直接影响y，那么z_{i10}由于不影响事前w，从而也不影响事前的y；如果有影响：z_{i10}不外生！

正式的识别假设

- 不失一般性用面板设定：

$$y_{it} = \beta w_{it} + x_{it}'\eta + u_{it}$$

其中x_{it}包含时间、个体固定效应等。

假设1

对于所有的s, m，回归：

$$w_{it} = x_{it}'\zeta + \sum_m \sum_s \delta_{ms} \times \mathbb{1}\{t = s\} \times z_{im0} + e_{it}$$

其中δ_{ms}有限且$\sum_s \sum_m g_{ms}\delta_{ms} \neq 0$。

假设2

对于所有的$g_m \neq 0$的m，有：$\mathbb{E}(u_{it}z_{im0}|x_{it}) = 0$。

打开Bartik工具的黑箱

- 根据以上介绍的等价关系，虽然B_{it}是一个工具变量，但是实际上等价于很多份额作为工具变量
  - 截面：份额作为工具变量
  - 面板：份额与时间固定效应的乘积作为工具变量
- 所以Bartik工具综合了很多很多工具变量的结果
  - 可以分解！
  - 可以直接使用这些多工具的结果进行估计，同时进行过度识别检验

Bartik工具变量估计的分解

Rotemberg权重（截面）

Bartik工具变量的结果可以分解为：

$$\hat{\beta}^{\text{Bartik}} = \sum_m \hat{\alpha}_m \hat{\beta}^m$$

其中$\hat{\beta}^m$为使用份额m作为工具的2SLS估计结果，$\hat{\alpha}_m$为Rotemberg权重，且

$$\sum_m \hat{\alpha}_m = 1$$

Bartik工具变量估计的分解

注意！

权重$\hat{\alpha}_m$有可能为负！从而：

- 如果同质性影响，$\hat{\beta}^m$都几乎相同，无影响
- 如果有异质性影响，$\hat{\beta}^{\text{Bartik}}$可能有问题
  - 使用bartik_weight（https://github.com/paulgp/bartik-weight）计算Rotemberg权重
  - 查看Rotemberg权重比较大的m的估计结果
  - 根据$\hat{\alpha}_m$的正负分组，并观察两个组别权重的和

Rotemberg权重

- 可以计算$\hat{\alpha}_m$与g_m的相关性：$\hat{\alpha}_m$多大程度上可以由g_m解释
- 可以计算$\hat{\alpha}_m$与z_{ik}的方差的相关性，其中的方差按照：

$$\frac{1}{N-1}\sum_i (z_{ik}-\bar{z}_k)^2$$

  计算
- 同时可以计算$\hat{\alpha}_m$与第一阶段F统计量$\hat{F}_k$的相关性：第一阶段F大的工具不一定权重大
- 对于面板，同理，不过由于此时工具变量有$(M-1)\times T$个，所以需要加总：

$$\hat{\alpha}_m=\sum_t \hat{\alpha}_{m,t}$$

Bartik工具所需要做的一些检验

为了保证Bartik工具能够正确识别，一系列检验需要进行：

- 检验1：检查份额z_{im0}与初始期的特征的相关性
  - 真正重要的是z_{im0}能否直接预测y（的变化、增长率等）
  - 如果z_{im0}与一些能够预测y变化的因素相关，可能会有遗漏变量的偏误
- 检验2：如果有政策变化——平行趋势检验
- 检验3：
  - 其他的IV估计量（LIML、MB2SLS、HFUL等）结果是否大致相同
  - 过度识别检验（直接使用份额做IV）
  - 异质性影响：同样会使得以上两个检验不通过：直接查看$\hat{\beta}^m$

Bartik工具：另一种视角

- 以上介绍Bartik工具聚焦份额的外生性，而Borusyak、Hull和Jaravel（2022）考虑了g_m的外生性
- 考虑模型

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

构建Bartik工具变量：

$$B_i = \sum_{m=1}^{M} z_{im} g_m$$

其中$i = 1, ..., N$为个体，$m = 1, ..., M$为部门。

- 这里需要假设$M\ (N) \to \infty$！
  - 需要大量的冲击g_m

识别讨论

- 根据以上设定，识别条件为

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

- 然而，注意到$\{B_i\}$不会是i.i.d的，类似的，$\{u_i\}$也有可能不是i.i.d的?
- 解决方法：转而使用矩条件

$$\mathbb{E}\left(\sum_{i=1}^{N} \omega_i B_i u_i\right) = 0$$

其中ω_i为回归权重，且$\sum_i w_i = 1$。

- 根据以上的识别条件，矩估计为

$$\hat{\beta} = \frac{\sum_{i=1}^{N} \omega_i B_i y_i^{\perp}}{\sum_{i=1}^{N} \omega_i B_i w_i^{\perp}}$$

识别条件

- 核心问题：以上识别条件何时成立?
- 将识别条件改写为

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{N}\omega_i B_i u_i\right) &= \mathbb{E}\left(\sum_{i=1}^{N}\omega_i \sum_{m=1}^{M} z_{im} g_m u_i\right) \\
&= \mathbb{E}\left(\sum_{m=1}^{M} g_m \sum_{i=1}^{N}\omega_i u_i z_{im}\right) \\
&= \mathbb{E}\left(\sum_{m=1}^{M} g_m \sum_{i=1}^{N}\omega_i z_{im} \frac{\sum_{i=1}^{N} u_i \omega_i z_{im}}{\sum_{i=1}^{N}\omega_i z_{im}}\right) \\
&\triangleq \mathbb{E}\left(\sum_{m=1}^{M} g_m s_m \bar{u}_m\right)
\end{aligned}
$$

识别条件

- 识别条件成为

$$\mathbb{E}\left(\sum_{m=1}^{M} s_m g_m \bar{u}_m\right) = 0$$

- 其中:
  - 新的权重

$$s_m = \sum_{i=1}^{N} \omega_i z_{im}$$

且$\sum_m s_m = \sum_m \sum_i \omega_i z_{im} = \sum_i \omega_i \sum_m z_{im} = \sum_i \omega_i = 1$。
  - 新的误差项

$$\bar{u}_m = \frac{\sum_{i=1}^{N} u_i \omega_i z_{im}}{\sum_{i=1}^{N} \omega_i z_{im}}$$

将地区层面误差项进行加权平均得到部门层面的平均误差项

估计方法

1. 使用y, w对x做回归得到残差$y^{\perp}, w^{\perp}$;
2. 计算暴露加权（exposure-weighted）平均

$$\bar{v}_m = \frac{\sum_{i=1}^{N} v_i \omega_i z_{im}}{\sum_{i=1}^{N} \omega_i z_{im}}, v \in \left\{y^{\perp}, w^{\perp}\right\}$$

3. 在回归

$$\bar{y}_m^{\perp} = \alpha + \beta \bar{w}_m^{\perp} + \bar{u}_m$$

中，使用g_m作为$\bar{w}_m^{\perp}$的工具变量

- 以上（冲击层面的）回归与使用ω_i加权的Bartik IV的估计完全等价（证明）
- 但是标准误不等价！

识别条件

- 问题：识别条件$\mathbb{E}\left(\sum_{m=1}^{M} s_m g_m \bar{u}_m\right) = 0$何时成立?
- 令$\bar{u} = \{\bar{u}_m, m = 1, ..., M\}, s = \{s_m, m = 1, ..., M\}$

第一组假设

1. （拟随机的冲击）$\mathbb{E}\left(g_n | \bar{u}, s\right) = \mu$
2. （不相关的冲击）赫芬达指数$\mathbb{E}\left(\sum_m s_m^2\right) \to 0$且$\mathbb{C}\left(g_m, g_{m'} | \bar{u}, s\right) = 0, m \neq m'$

识别条件

- 如果假设$\mathbb{E}\left(g_n|\bar{u},s\right)=\mu$，那么

$$
\begin{aligned}
\mathbb{E}\left(\sum_{m=1}^{M} s_m g_m \bar{u}_m\right) &= \mathbb{E}\left[\mathbb{E}\left(\sum_{m=1}^{M} s_m g_m \bar{u}_m \middle| \bar{u}, s\right)\right] \\
&= \mathbb{E}\left[\sum_{m=1}^{M} s_m \bar{u}_m \mathbb{E}\left(g_m \middle| \bar{u}, s\right)\right] \\
&= \mu\mathbb{E}\left(\sum_{m=1}^{M} s_m \bar{u}_m\right) = \mu\mathbb{E}\left(\sum_{m=1}^{M}\sum_{i=1}^{N} u_i \omega_i z_{im}\right) \\
&= \mu\mathbb{E}\left(\sum_{i=1}^{N} u_i \omega_i \sum_{m=1}^{M} z_{im}\right) = \mu\mathbb{E}\left(\sum_{i=1}^{N} u_i \omega_i\right) \\
&= 0
\end{aligned}
$$

识别条件

- 令q_m为部门的控制变量，令$q = \{q_m, m = 1, ..., M\}$
  - 当然也可以是部门的分类$c(m) \in \{1, ..., C\}$

第二组假设

1. （条件拟随机的冲击）$\mathbb{E}(g_n | \bar{u}, s, q) = q_m' \mu$
2. （不相关的冲击残差）赫芬达指数$\mathbb{E}\left(\sum_m s_m^2\right) \to 0$且$\mathbb{C}(g_m, g_{m'} | \bar{u}, q, s) = 0, m \neq m'$

- 如果使用Bartik IV的估计方法，需要额外控制$q_i = \sum_m z_{im} q_m$
- 使用冲击层面的IV估计只需要加入q_m作为控制变量

识别条件

- 识别条件中冲击不相关的假定依然很强，两个放松的条件：

不相关假设的两种放松

- 存在部门的一个划分，使得$\mathbb{E}\left(\sum_c s_c^2\right) \to 0$，其中$s_c = \sum_{m:c(m)=c} s_m$且$\mathbb{C}\left(g_m, g_{m'} | \bar{u}, q, s\right) = 0, c\left(m\right) \neq c\left(m'\right)$
- 存在一个序列$B_M \geq 0$以及一个函数$f\left(\cdot\right)$满足$\sum_{r=1}^{\infty} f\left(r\right) < \infty$，使得$B_M \mathbb{E}\left(\sum_m s_m^2\right) \to 0$且$|\mathbb{C}\left(g_m, g_{m'} | \bar{u}, q, s\right)| \leq B_M \cdot f\left(|m' - m|\right)$

估计问题

- 然而标准误可能有问题（Adao, Kolesar和Morales，2019）
  - 直接使用冲击层面的IV估计，并计算稳健标准误
  - Stata：ssaggregate

合成工具变量

- 工具变量（解释变量）需要满足外生性，比如一些自然实验的冲击
- 然而有时，有些冲击不得不与一些内生的变量合在一起才能被观察到：
  - Bartik工具：可能g_m是外生的，但是z_{im}的内生的
  - 交通设施：哪些地方开通高铁可能不是外生的，但是开通高铁的时间可能是外生的
  - 美国每个州的（给穷人的）健康保险政策是外生的，但是人口特征不是（比如可能跟健康状况有关）
- 问题：这些工具变量是否可以使用？如何使用？Borusyak和Hull（2021）回答了这一问题。

一个示例：市场进入

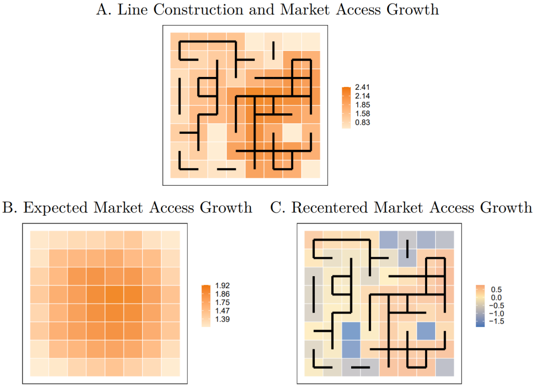
- 考虑交通基础设施（市场进入，MA_i）对土地价格V_i的影响：

$$\Delta \ln V_i = \beta_0 + \beta \Delta \ln MA_i + u_i$$

- 假设在一个棋盘上随机取点修建铁路
- 虽然哪个个点修建铁路是外生的，但是地理位置是内生的！
  - 棋盘中间天然具有更高的市场进入
- 解决办法：
  - 如果棋盘上的点是随机被选取出来修建铁路的，计算一个期望的市场进入$\mu_i = \mathbb{E}(\Delta \ln MA_i)$
  - 使用$\Delta \ln MA_i - \mu_i$作为解释变量
  - μ_i：可以通过模拟计算得到

市场进入示例

Figure 1: Market Access Growth in the Motivating Example



设定

- 假设：

$$z_i = f_i\left(g, x\right)$$

  其中：
  - $f_i\left(\cdot\right), i = 1, ..., N$为已知函数
  - g为$N \times 1$的外生冲击
  - x为一些可观测变量
- 回归方程：

$$y_i = \beta_0 + \beta w_i + u_i$$

重要引理

假设1

假设冲击是外生的，即$g \perp\!\!\!\perp u|x$

在以上条件下，由于：

$$
\begin{aligned}
\mathbb{E}\left(\frac{1}{N}\sum_i z_i u_i\right) &= \mathbb{E}\left(\frac{1}{N}\sum_i f_i\left(g,x\right)u_i\right) \\
&= \mathbb{E}\left(\mathbb{E}\left[\frac{1}{N}\sum_i f_i\left(g,x\right)u_i|x\right]\right) \\
&= \mathbb{E}\left(\frac{1}{N}\sum_i u_i \mathbb{E}\left[f_i\left(g,x\right)|x\right]\right) \\
&= \mathbb{E}\left(\frac{1}{N}\sum_i u_i \mu_i\right)
\end{aligned}
$$

其中$\mu_i = \mathbb{E}\left[f_i\left(g,x\right)|x\right]$为工具变量的期望。

新的工具

- 根据以上结论，有：

$$\mathbb{E}\left(\frac{1}{N}\sum_{i}\left(z_i-\mu_i\right)u_i\right)=0$$

- 两种做法：
  - 计算得到μ_i使用$z_i-\mu_i$作为工具变量：重新中心化（recentering）
  - 直接将μ_i作为控制变量、z_i作为工具变量
    - 启示：如果μ_i被其他控制变量、固定效应吸收了，不用做额外处理，使用z_i做工具就够了。

工具期望的计算

问题是：如何计算μ_i?

- 假设我们知道给定x，g的分布$G(g|x)$，那么可以直接计算：

$$\mu_i = \int f_i(g; x)\, dG(g|x)$$

- 然而$G(\cdot|\cdot)$可能不好设定：直接将g向量进行重新排列组合，得到新的组合$\pi(g)$，然后计算$f_i(\pi(g); x)$，重复多次后计算均值即可。
  - 比如对于高铁开通，如果使用2016年的数据，假设开通时间是随机的，那么将2016已经开通的城市和有规划的城市进行一个排列组合，取MA的均值即可。
  - 对于保险资格，假设某个个体随机属于某一个州，按照这个州的标准判断其在这个州是不是有资格，然后求平均