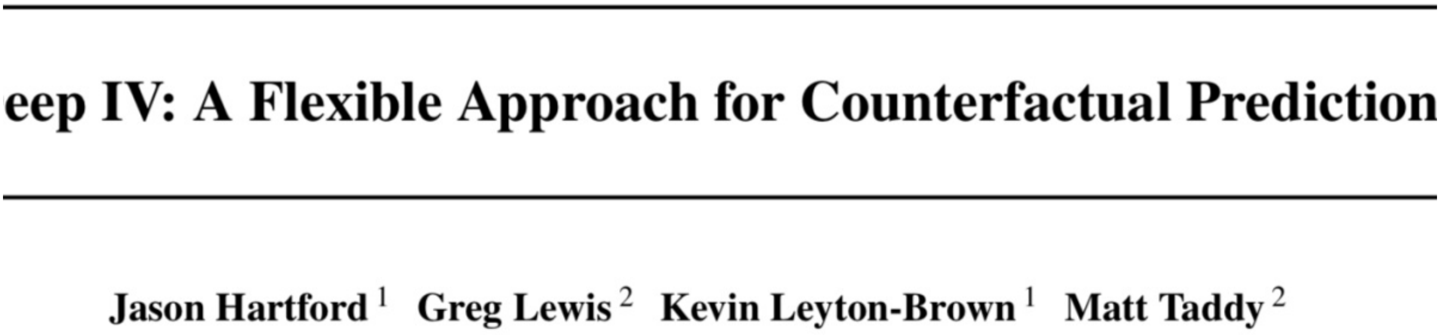

eep IV: A Flexible Approach for Counterfactual Prediction

Jason Hartford [1] Greg Lewis [2] Kevin Leyton-Brown [1] Matt Taddy [2]

【统计】Instrumental Variables

张楚珩
清华大学 交叉信息院博士在读

还是继续前面因果推断的话题，这里讲一个从数据中做因果推断的另一种有效的方法——Instrumental variables（IV）。同时，也补充一个用深度学习方法来实现 IV 的一篇论文（ICML 2017）。

原文传送门

soderbom.net/lec2n_fina...

http://www3.grips.ac.jp/~yamanota/

Hartford, Jason, et al. "Deep IV: A flexible approach for counterfactual prediction." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

正文

一、Instrumental variable

还是看一个线性模型，

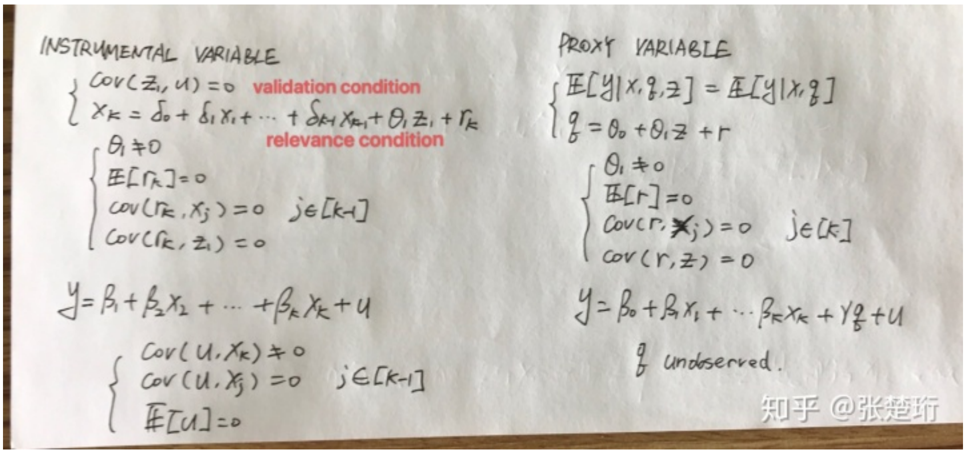
$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_K x_K + u, \qquad (2.1)$$

前面说到，线性模型前面的系数表征了自变量对于因变量的因果关系，一个线性模型要能使用 OLS 估计系数需要满足 exogeneity 的要求。现在假设其中某一个变量不满足该要求，即

$\mathbb{E}[u] = 0, Cov(x_j, u) = 0, j \in [K-1], Cov(x_K, y) \neq 0$

这时，我们引入 instrumental variable 方法来解决内生性问题，并且为 β_j 的提供一致性估计。

一个 instrumental variable z_1 需要满足的性质（和它与 proxy variable 的对比，一块放上来）



validity condition and relevance condition （图中写错了）

第一项的要求应该两个是等价的，即额外选择的这个变量不能成为解释变量进入原来的方程中。最主要的区别在第二个要求上，proxy variable 要求找到的变量能够解释所有 q 和其他 解释变量的关联（即，q 剔除掉 z 的影响之后，r 和其他解释变量无关）；而 instrumental variable 则希望找到一个和其他变量都没啥关系，但是只和 x_K 有一点关联的变量。

论坛 (StackExchange)上说：instrumental variable 的目的是想找因果关系，减小 estimation error 产生的影响，对于它来说， x_K 能观察到，但是可能有偏差，因此要找一个只影响 x_K 的变量来抵消相应的估计误差；proxy variable 是想想办法把原来的线性模型系数估计处理，其中的变量 q 观察不到，想要找一个和它接近的变量来替换它。

二、instrumental variable estimator

令

$$\begin{aligned}\mathbf{x} &= \begin{bmatrix} 1 & x_2 & x_3 & \dots & x_K \end{bmatrix} \\ \mathbf{z} &= \begin{bmatrix} 1 & x_2 & \dots & x_{K-1} & z_1 \end{bmatrix},\end{aligned}$$

假设样本数目为 N，那么它们都是 NxK 的矩阵。

方程可以写为

$$y = \mathbf{x}\boldsymbol{\beta} + u,$$

其中 y 和 u 都是 Nx1 的向量， β 是 Kx1 的向量。

根据前面的条件，有

$$E\left(\mathbf{z}'u\right)=\mathbf{0}.$$

不难推得

$$\begin{aligned} E\left(\mathbf{z}'\left(y-\mathbf{x}\boldsymbol{\beta}\right)\right) &= \mathbf{0} \\ E\left(\mathbf{z}'\mathbf{x}\right)\boldsymbol{\beta} &= E\left(\mathbf{z}'y\right), \end{aligned}$$

由于 $\theta_1 \neq 0$ ，即 x_K, z_1 相互关联，因此

$$\operatorname{rank} E\left(\boldsymbol{z}'\boldsymbol{x}\right)=K,$$

这样我们可以对其求逆，解出系数

$$\boldsymbol{\beta}=\left[E\left(\mathbf{z}'\mathbf{x}\right)\right]^{-1}E\left(\mathbf{z}'y\right).$$

系数的估计公式为

$$\hat{\boldsymbol{\beta}}^{IV}=\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'Y\right), \tag{3.2}$$

称它为 **instrumental variable estimator**。

可以验证，它为一致性估计

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{IV} &= \left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'\left(\mathbf{X}\boldsymbol{\beta}+u\right)\right) \\ &= \boldsymbol{\beta}+\left(\mathbf{Z}'\mathbf{X}\right)^{-1}\left(\mathbf{Z}'u\right). \end{aligned}$$

Using Slutsky's theorem, we get

$$\begin{aligned} p\lim\hat{\boldsymbol{\beta}}^{IV} &= \boldsymbol{\beta}+\left[E\left(\mathbf{Z}'\mathbf{X}\right)\right]^{-1}E\left(\mathbf{Z}'u\right) \\ &= \boldsymbol{\beta}, \end{aligned}$$

三、Two-Stage Least Squares

前面只使用一个 Instrumental variable，也可以使用多个 Instrumental variable；对于多个 instrumental variable 的情形，这里使用 two-stage least square （2SLS）方法来解决。

Instrumental variable 的要求

对于多个 instrumental variable，前面对于 instrumental variable 的两个要求（validity、relevance）可以写为

$$E(\boldsymbol{z}'u) = 0 \qquad \text{(Validity)}$$

$$\operatorname{rank}(\boldsymbol{z}'\boldsymbol{x}) = K, \qquad \text{(Relevance)}$$

其中 z 的维度为 1xL，它包含所有 exogenous explanatory variables；同时要求 $\operatorname{rank}(z'z) = L$ ，即 z 中的各个元素之间没有共线性（collinearity）。K 表示 explanatory variable 的个数。

具体来说，它首先要求 instrumental variable 的数目要至少和 endogenous explanatory variable 的数目一样多；同时 instrumental variable 和 endogenous variable 之间要有比较好的对应关系。比如对于 x_3, x_4 两个 endogenous variable 和 z_1, z_2 两个 instrumental variable

$$\begin{aligned} x_3 &= \pi_1 + \pi_2 x_2 + \pi_3 z_1 + \pi_4 z_2 + \varepsilon_1, \\ x_4 &= \gamma_1 + \gamma_2 x_2 + \gamma_3 z_1 + \gamma_4 z_2 + \varepsilon_2. \end{aligned}$$

如果 $\pi_3 = \gamma_3 = 0, \pi_4 \neq 0, \gamma_4 \neq 0$ 就不行，因为 z_1 与 x_3, x_4 都不相关了。

2SLS 方法

考虑一个线性模型

$$y_1 = \alpha_1 y_2 + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u \qquad (1)$$

其中 $x_1, \cdots, x_k$ 为 exogenous variables， y_2 为 endogenous variable。令 $\mathbf{z} = (1, x_1, \cdots, x_k, z_1, \cdots, z_m)$ ，其中最后 m 个元素为 m 个 Instrumental variables。 y_2 可以被写作

$$\begin{aligned} y_2 &= \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + \delta_{k+1} z_1 + \ldots + \delta_{k+m} z_m + \varepsilon \\ &= \hat{y}_2 + \varepsilon \end{aligned}$$

可以看出 $cov(y_2, u) = cov(\varepsilon, u) \neq 0, cov(\hat{y}_2, u) = 0$ ，即可以用 $\hat{y}_2$ 代替 y_2 用于线性回归中然后再用 OLS 估计系数 $\alpha_1, \beta_{0:k}$ 。

y_2 对 $\mathbf{z} = (1, x_1, \cdots, x_k, z_1, \cdots, z_m)$ 做回归，可以得到其预测值

$$\hat{y}_2 = Z\hat{\delta} = Z(Z'Z)^{-1} Z' y_2$$

这样的做法等价于一个 Nx(k+1) 的矩阵 $X=(y_2,x_1,\cdots,x_k)$ 对 $\mathbf{z}=(1,x_1,\cdots,x_k,z_1,\cdots,z_m)$ 做回归，其预测值为

$$\hat{X} = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X = P_Z X$$

其中 $\hat{\Pi}$ 为一个 (k+m+1)x(k+1) 的矩阵，其看起来为

$$\hat{\Pi} = \begin{bmatrix} \delta_1 & 1\ 0 & 0 \\ \delta_2 & 0\ 1 & 0 \\ & & \\ \delta_{k+1} & 0\ 0 & 1 \\ \vdots & 0\ 0 & 0 \\ \delta_{k+m+1} & 0\ 0 & 0 \end{bmatrix}.$$

即，其他的 $x_1,\cdots,x_k$ 还是原来的数值，只是把 y_2 换成了 $\hat{y}_2$ 。仿照 IV estimator

$$\hat{\beta}_{IV} = (Z'X)^{-1}Z'Y$$

把 z 替换文 $\hat{x}$ ，可以得到

$$\begin{aligned} \hat{\beta}_{2SLS} &= (\hat{X}'X)^{-1}\hat{X}'Y \\ &= (X'P_Z X)^{-1}X'P_Z Y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'Y \end{aligned}$$

即，在估计的 variables $\hat{x}$ 上做 OLS 即可。

四、 一个例子

DeepIV 可以看做把上述 2SLS 方法推广到非线性模型。在讲这篇 paper 之前，先讲一个这篇 paper 提到的一个例子，用来理解 endogeneity。

Figure 1. (Left) Our air-travel demand example, with arrows representing causal effects. Price is the policy variable, sales is the outcome, and holidays are observable covariates. There is a big 'conference', unobserved to the policy-maker, that drives demand and (due to the airline's pricing algorithms) price. The instrument is the cost of fuel, which influences sales only via price. (Right) The general structure of the DGP under our IV specification; x represents observable features, p is our treatment variable of interest, z represents the instruments, and latent effects e influence the outcome y additively.

考虑研究航空公司票价（prince，p）和销量（sales，y）的影响，即当其他条件不变时，如果增加票价，销量会如何变化；这是在研究一个 causal effect。

最好的方法就是随机化地做实验，比如随机对一些顾客，提高或者降低面向他们的票价；但是这样的实验不仅会对航空公司造成实际的经济损失（损失了潜在的客户或者减少了盈利），而且还客户还因为不公平的对待而对航空公司不满。

因此，我们希望从历史数据里面来挖掘票价（P）和销量（Y）之间的因果关系。

1. Confounders：一个直接的方法就是把 Y 对 P 做回归，但是这样往往得出错误的结论。比如可能存在另外的因素，节假日（X），它不仅影响票价（P）也同时影响销量（Y），在节假日的时候航空公司定价高（P 大），同时人们出行需求也大，因此销量也高（Y 大）；这是在历史数据里面做回归，得出的结论就会使 Y 和 P 成正相关，这显然是错误的。
2. Unobservable variables：有一种方法是把可以观察到的影响因素也放到回归方程里面，根据前面几个 post 的内容，如果能够把所有的因素都包含进来，那么也能够得出正确的结论。但是，有可能会有一些观察不到的作用因素（E），比如在某个地区召开某个会议，它会同时使得票价（P）和销量（Y）都变大。
3. Instrumental variable（IV）：因此，最有效的办法就是找到一个 IV，使得它只影响票价（当然，它也可以通过 P 来影响 Y，但是不能直接影响 Y），而和其他因素（X、E）不相关。在这里，我们找到『燃油价格』这个因素来做为 IV，它上升时，航空公司基于成本考虑会上调票价，但是该因素和其他因素（比如是否节假日、是否在某个地区有会议）无关。需要注意的是，IV 的选取基本上基于人为的先验知识。

五、问题设定

符号和前面一致，我们想要探究的是 y 和 p 之间的关系，模型为

$$y = g(p, x) + e. \qquad (1)$$

其中 $\mathbb{E}[e]=0$ ，认为变量 p 是 endogenous 的，即 $\mathbb{E}[e|x,p]\neq 0$ ，或者 $\mathbb{E}[pe|x]\neq 0$ （是前面的一种特殊情况，相当于线性不相关）。

定义一个 counterfactual prediction function

$$h(p, x) \equiv g(p, x) + \mathbb{E}[e|x], \qquad (2)$$

注意到，要是 e 和 x 无关，而只与 p 有关，那么 h 函数和 g 函数一致。固定 x，希望知道如果把 $p=p_0$ 变动为 $p=p_1$ 时 y 会变化多少，这其实就是在求 $g(p_1,x)-g(p_0,x)=h(p_1,x)-h(p_0,x)$ 。

这里再次强调一下，函数 $h(p,x)$ 不能通过用 p, x 去拟合 y 来得到，因为拟合得到的函数实际上为 $\mathbb{E}[y|p,x]$ ，它们的差为

$$\begin{aligned}\mathbb{E}[y|p_1,x]-\mathbb{E}[y|p_0,x]\\ = g(p_1,x)-g(p_0,x)+\Big(\mathbb{E}[e|p_1,x]-\mathbb{E}[e|p_0,x]\Big).\end{aligned}$$

不等于我们需要求解的 $g(p_1,x)-g(p_0,x)$ 。

这时，我们就需要一个 Instrumental variable z 了。它需要满足以下三点要求

Relevance $F(p|x,z)$, the distribution of p given x and z, is not constant in z.

Exclusion z does not enter Eq. (1)—i.e., $z \perp\!\!\!\perp y \mid (x,p,e)$.

Unconfounded Instrument z is conditionally independent of the error—i.e., $z \perp\!\!\!\perp e \mid x$.[3]

第一点要求和前面的 $\theta_1 \neq 0$ 类似，只不过变成了在非线性模型下的表述；第二个即表明引入的该变量不会成为 explanatory variable 进入原方程；第三个表示引入的该变量和原来的误差项不相关。

六、DeepIV 结构

在引入变量 z 之后，可以写出如下关系

$$\begin{aligned}\mathbb{E}[y|x,z] &= \mathbb{E}[\,g(p,x)|x,z]+\mathbb{E}[e|x]\\ &= \int h(p,x)dF(p|x,z), \qquad (5)\end{aligned}$$

注意到 $\mathbb{E}[y|x,z]$ 可以通过有监督学习得到， $F(p|x,z)$ 也可以通过有监督学习得到，在得到这两者之后，能够反向求解得到我们需要的 $h(p,x)$ 。实际算法分为两步：

- 第一步：先学习 $F_\phi(p|x,z)$ ；对于离散的 p，学习一个 categorical distribution， $P(p=p^k)=\pi_k(x,z;\phi)$ ；对于连续的 p，学习一个 mixture of Gaussian distributions， $y|x,z\sim\sum_k \pi_k(x,z;\phi)\mathcal{N}(\mu_k(x,z;\phi),\sigma_k(x,z;\phi))$ 。
- 第二步：然后学习 $h_\theta(p,x)$ ，目标是最小化

$$\mathcal{L}(D;\theta)=|D|^{-1}\sum_t\left(y_t-\int h_\theta(p,x_t)d\hat{F}_\phi(p|x_t,z_t)\right)^2. \quad (7)$$

其中 |D| 是训练集的大小。

对于第二步，文章着重讲了一下优化的方法，令 $\mathcal{L}_t=\left(y_t-\mathbb{E}_{p\sim F_\phi(x_t,z_t)}[h_\theta(p,x_t)]\right)^2$ ，而

$$\mathbb{E}_{\mathcal{D}}\nabla_\theta \mathcal{L}_t = -2\mathbb{E}_{\mathcal{D}}\Big(\mathbb{E}_{F_\phi(p|x_t,z_t)}\big[y_t - h_\theta(p^k, x_t)\big] \cdot \mathbb{E}_{F_\phi(p|x_t,z_t)}\big[h'_\theta(p^k, x_t)\big]\Big) \quad (9)$$
$$\neq -2\mathbb{E}_{\mathcal{D}}\mathbb{E}_{F_\phi(p|x_t,z_t)}\big[\big(y_t - h_\theta(p^k, x_t)\big)\, h'_\theta(p^k, x_t)\big],$$

因此可以看出，在对其估计梯度的时候，对于每一个 (x_t, z_t, y_t) 数据点需要从分布 $F_\phi(x_t, z_t)$ 从采集两波 p 样本，即

$$\widehat{\nabla}_\theta^B \mathcal{L}_t \equiv -2\left(y_t - B^{-1}\sum_b h_\theta(\dot{p}_b, x_t)\right)B^{-1}\sum_b h'_\theta(\ddot{p}_b, x_t). \quad (10)$$

因此要过两波 forward pass 和一波 backward pass（求梯度）。

如果省事，只采样一波 p，在两处地方都用它们，那么得到的损失函数将会是原损失函数的上界。

$$\hat{\mathcal{L}}(D;\theta) \leq |D|^{-1}\sum_t \sum_{\dot{p}\sim\hat{F}_\phi(p|x_t,z_t)} \left(y_t - h_\theta(\dot{p}, x_t)\right)^2. \quad (11)$$

但是实验中发现这样做效果竟然更好，文章给出的解释是，这样的梯度下降虽然有 bias，但是更新的 variance 更小。

这里讨论的情形是 continuous outcome space and continuous treatment space，此外还可以推广到 discrete outcome space / treatment space。

对于因果推断来说，模型的 validation 更为重要，文章还给出了做 validation 的方法。