【统计】Endogeneity Bias

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，不过这里更多地来讨论线性模型中的内生性误差（endogeneity bias），一个更通俗的名字叫做 sample selection bias。仍然基于一个网上的 lecture notes。

原文传送门

soderbom.net/lec1n_fina...

摘要

1. 在一定条件下，线性回归的系数表征了变量之间的因果关系；
2. 在没有内生性误差的情形下，可以通过 ordinary least square（OLS）可以得到对线性回归系数的一致性估计；
3. 介绍了三种可能产生内生性误差的情况，并针对一种情况给出了解决方法；
4. 对于其他情况的解决方法将在后面讲。

正文

一、Endogeneity Bias

回顾一下前一讲的内容，要做因果推断，最为有效的方法是做试验。但是实际情况中，还是只能从数据里面挖掘因果关系（虽然 notes 里面讲了，这样做非常危险）。从观测数据里面挖掘因果关系可能最大的问题就是有一些潜在的因素 C 会同时影响要研究的 W 和 Y，可以通过把这些因素包含和考虑尽量从而消除它们的影响（adjusting for confounders）。

可以用 partial effect 来衡量 W 对 Y 的影响，即如果保持其他的因素 C 不变，改变 W 会如何地影响 Y。它被定义为偏导数，如果 X 变量为二值的，那么它就是前一讲里面的 θ。

If w is continuous, the partial effect is

$$\frac{\partial E(y|w,c)}{\partial w},$$

while if w is a dummy variable, we would look at

$$\text{theta} = E(y|w=1,c) - E(y|w=0,c).$$

当然，还有其他的衡量方法，比如 elasticity 和 semi-elasticity 等，这里不再贴出来。

线性模型

特别地，前面提到，如果所有的 confounders Z 都被包含进来了，并且线性模型足够正确，那么把 Y 相对于 X 和 Z 做线性回归，X 前面的系数就表征了 X->Y 之间的因果关系。因此，我们希望能够对于线性模型做回归从而得到变量之间的关系。

那么能不能对于线性模型做回归，得到有效的系数呢？考虑一个线性模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + u,$$

考虑通过 ordinary least-square （OLS）回归估计系数 β 。一个好的估计方法应该具有一致性（consistency），即当样本数量足够多的时候，它能够估计的无限准。要想能通过 OLS 一致估计系数，需要满足以下两个要求，即 1）噪声 u 需要是零均值；2）噪声 u 需要和前面的所有 explanatory variable （即这里的 $x_1,\cdots,x_K$ ）都不相关。

$$E(u) = 0,$$

$$Cov(x_j, u) = 0, \quad j = 0, 1, ..., K. \tag{4.1}$$

第一点要求是很好满足的，因为可以通过常数项系数 β_0 来消去噪声 u 的均值部分；第二点要求是比较难满足的，如果某一个变量 x 和噪声相关，那么称变量 x 是 endogenous 的（中文是『内生性』的意思，虽然理解起来仍然比较晦涩），也称这样的线性回归有 endogeneity bias。

为了理解噪声 u 为什么需要和 explanatory variables 不相关，我们考虑一个最简单的线性模型

$$y = \beta_0 + \beta_1 x_1 + u.$$

前面说到均值部分不是很关键，因此做变量代换 $\tilde{y} = y - \bar{y}$ ， $\tilde{x}_1 = x_1 - \bar{x}_1$ 消去常数项系数

$$\tilde{y} = \beta_1 \tilde{x}_1 + u,$$

根据 OLS estimator 的公式（可以看维基Wiki: OLS Simple_linear_regression_model）有，

$$\hat{\beta}_1^{OLS} = \beta_1 + \frac{\sum_i \tilde{x}_{1i} u_i}{\sum_i \tilde{x}_{1i}^2},$$

注意到，当 $Cov(x_1, u) \neq 0$ 时，有

$$\begin{aligned} p\lim \hat{\beta}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i} u_i}{\sum_i \tilde{x}_{1i}^2}, \\ &= \beta_1 + \frac{Cov(x_1, u_i)}{var(x_1)} \neq \beta_1, \end{aligned} \tag{4.2}$$

其中 $p\lim$ 代表样本数目足够多时的极限。

结论

因此，从这一个部分可以看出，我们可以通过做线性回归来估计 partial effect（其实也是一种因果关系）；但是要能这么做需要满足两点重要的要求 1）模型本身是线性的；2）不能有 endogeneity bias。

二、引起 endogeneity bias 的原因

Notes 里面提到引起 endogeneity bias 的原因主要有一下三点：

- Omitted variables
- Measurement error
- Simultaneity

Omitted Variables （omitted variable bias）

考虑一个线性模型

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,$$

这个线性模型满足前面提到的要求

$$Cov\left(x_1, u\right) = Cov\left(x_2, u\right) = 0.$$

即，如果能观察到数据集 $\{(x_{1i}, x_{2i}, y_i)\}$ ，那么对它做 OLS 就能一致性地估计到相应的系数。

但是现实中存在的问题是，可能有一个变量（比如 x_2 ）我们没有办法观察到，那么我们还能够准确估计到另一个变量 x_1 前面的系数么？

这时，我们要估计的模型变为

$$y = \gamma_0 + \gamma_1 x_1 + \varepsilon_i,$$

it must be that $\varepsilon_i = (\beta_2 x_{2i} + u_i)$.

用 OLS 估计 γ_1 ，有

$$\begin{aligned} p\lim \hat{\gamma}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i}\left(\beta_2 \tilde{x}_{2i} + u_i\right)}{\sum_i \tilde{x}_{1i}^2}, \\ &= \beta_1 + \beta_2 \frac{Cov\left(x_1, x_2\right)}{var\left(x_1\right)}, \end{aligned}$$

因此不难看出，一般情况下，缺少变量之后 $\gamma_1 \neq \beta_1$ ，除非 1）本身缺少的变量对 Y 就没有贡献，即 $\beta_2 = 0$ 或者 2）缺少的变量和要研究的变量之间没有关联，即 $Cov(x_1, x_2) = 0$ 。

Measurement error （attenuation bias）

考虑一个单变量线性回归

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i.$$

如果对于 Y 的测量有误差，而该误差独立，并且均值为零；那么它就能够被放入噪声项 u 中，对于系数的估计不产生影响。

如果对于 X 的测量，同时该误差也独立，均值也为零会怎样呢？考虑 X 的观测值含有误差

$$x_{1i}^{obs} = x_{1i} + v_i,$$

那么我们估计的方程实际上是

$$y_i = \beta_0 + \beta_1 x_{1i}^{obs} + e_i, \tag{5.1}$$

where $e_i = (u_i - \beta_1 v_i)$.

注意到虽然误差项 e 中的 u 和 v 都是独立的，但是这里面包含了一个系数 β_1 。这时，对于 β_1 的 OLS 估计为

$$\begin{aligned}
p\lim \hat{\beta}_1^{OLS} &= \beta_1 + p\lim \frac{\sum_i \tilde{x}_{1i}^{obs} e_i}{\sum_i \left(\tilde{x}_{1i}^{obs}\right)^2}, \\
&= \beta_1 + p\lim \frac{\sum_i \left(\tilde{x}_{1i} + v_i\right)\left(u_i - \beta_1 v_i\right)}{\sum_i \left(\tilde{x}_{1i} + v_i\right)^2}, \\
&= \beta_1 + \frac{-\beta_1 \sigma_v^2}{\sigma_{\tilde{x}_1}^2 + \sigma_v^2}, \\
&= \beta_1 \left(\frac{\sigma_{x_1}^2}{\sigma_{x_1}^2 + \sigma_v^2}\right)
\end{aligned}$$

注意到，

- 有测量误差的时候，估计到的系数在绝对值上都会减小，因此，这样的误差也叫做 attenuation bias。
- 偏离的程度和 $\sigma_v^2/\sigma_{\tilde{x}}^2$ 有关，这个比率就是信噪比。
- 虽然会产生 bias，但是其符号不变。

Simultaneity

有的时候 explanatory variables （X）会和 dependent variables （Y）混到一起，比如

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + u_1, \tag{5.2}$$

$$y_2 = \beta_0 + \beta_1 y_1 + \beta_2 x_2 + u_2, \tag{5.3}$$

注意到，这样写出来的含义是对于单个的方程来讲， u_1, u_2 都满足前面 exogeneity 的要求；如果要对 y_1 在 x_1, x_2 上做回归，不难看出

$$y_1 = \frac{1}{\alpha_1\beta_1 - 1}[(\alpha_0 + \alpha_1\beta_0) + \alpha_2 x_1 + \alpha_1\beta_1 x_2 + (u_1 + \alpha_1 u_2)]$$

关键问题还是跟起那么一样，噪声项里面含有了要估计的系数，因此，OLS 得到的系数和原本的系数不一致。

三、Proxy Variable

前面讲到了 endogeneity bias 产生的三个原因，这里后面会讲两个解决方法：针对 omitted variable bias，可以使用 proxy variable 来解决；针对某一个 endogenous 变量，可以使用 instrumental variable 来解决。这里先讲前一个，后面方法会单独开一个 post 来讲。

考虑如下线性模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \gamma q + u, \tag{6.1}$$

其中变量 q 不能被观察到，这时候我们考虑能不能使用一个另外的变量 z 来代替它；或者说，另外找到的一个变量需要满足怎样的性质，才能使得 OLS 回归得到的系数不受 q 观察不到的影响？

总体来说，我们找到的这个变量 z 需要满足以下两个条件

1. 变量 z 必须和 q 相关，并且要能解释所有 q 和 x 们的关联性；
2. 如果给定了 x 们和 q，那么不管 z 取什么值，都不会改变 y，即 z 只通过影响 q 来影响 y。

数学上来说，考虑

$$q = \theta_0 + \theta_1 z + r,$$

需要满足

$$\begin{aligned} E(r) &= 0 \text{ (holds by def.)} \\ Cov(z, r) &= 0 \text{ (holds by def.)} \\ \theta_1 &\neq 0 \text{ (if not, useless proxy)} \\ Cov(x_j, r) &= 0 \text{ (crucial!).} \end{aligned}$$

以及

$$E(y|\mathbf{x}, q, z) = E(y|\mathbf{x}, q).$$

从这些条件出发，能推出 OLS 能对于系数做出一致性估计，

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \gamma\left(\theta_0 + \theta_1 z + r\right) + u \\ &= \left(\beta_0 + \theta_0\right) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \gamma\theta_1 z + \left(\gamma r + u\right). \end{aligned}$$

（后面考虑用多元 OLS 的公式 $\hat{\beta} = (X^T X)^{-1}(X^T Y)$ 再利用 Slutsky's theorem 就能得到）

编辑于 2019-10-25

回归分析　计量经济学　统计学

文章被以下专栏收录

强化学习前沿
读呀读paper

进入专栏