内生性和因果关系

李德山

[1] 西南科技大学 经济管理学院

[2]School of Economics and Management
Southwest University of Science and Technology

2021 年 3 月 18 日

主要内容

1. 什么是内生性问题？

主要内容

1. 什么是内生性问题？
2. 内生性的来源

主要内容

1. 什么是内生性问题？
2. 内生性的来源
3. “合格”的控制变量

什么是内生性问题

- **内部和外部有效性**概念提供了评价计量经济学研究是否准确回答了我们感兴趣的具体问题的一般框架。
  - 如果有关因果效应的统计推断对研究总体是正确，则称该统计分析是内部有效的。
  - 如果从研究总体及其环境中得到的相关推断和结论可推广到其他总体及其环境中，则称该统计分析是外部有效的。
  - 这里的“环境”指制度、法律、社会和经济环境。

什么是内生性问题

- 内部有效性由两部分组成：
  - 1. 因果效应估计量是无偏的、一致的。
  - 2. 假设检验具有要求的显著水平，且置信区间也具有要求的置信水平。
- 外部有效性的潜在威胁来自研究总体及其环境与感兴趣的总体和环境之间的差异。
  - 总体差异。例如，对小白鼠的起作用的药物对人类也能起作用吗？毕竟小白鼠总体与人类总体存在非常大的差异。
  - 环境差异。即使所研究的总体和感兴趣的总体相关，只要环境存在差异，一个研究的结论也可能不能推广到更一般。例如，美国的小班教学模式，是否对中国也有用？

什么是内生性问题

- 如何评估研究的外部有效性？
  - 一般，两个或两个以上研究得到的结论相似则支持外部有效性。
  - 但如得到了不能解释的结论差异，我们应该怀疑其外部有效性。
- 当利用回归模型进行预测时，对外部性有效性的关注就显得非常重要，而对因果效应的无偏估计关注就不那么重要了。

什么是内生性问题

- 因果问题涉及两个理论性概念之间的关系：原因和结果。$X \to Y$

例子：医疗的效果

通过观察法得到的医院效果 = 那些去了医院的人的健康状况 − 那些没去医院的人的健康状况

什么是内生性问题

- 但是，这两群人也许是不同的，去医院的人群健康状况往往较差。因此，通过观察获得的效果可能不是真实效果！

- 因果推断的根本问题：没有足够的“反事实”。没法避免，只能减少推论的不确定性。

什么是内生性问题

- 但是，这两群人也许是不同的，去医院的人群健康状况往往较差。因此，通过观察获得的效果可能不是真实效果！

- 因果推断的根本问题：没有足够的“反事实”。没法避免，只能减少推论的不确定性。

- 我们用数学公式来重新表达一下基于观察法的医院效果：

$$\overbrace{E(Y_{1i}|D_i=1)}^{\text{该去医院也的确去了的那些人}} - \overbrace{E(Y_{0i}|D_i=0)}^{\text{不该去医院的人且实际上也没去的那些人}} \tag{1}$$

$$\underbrace{\overbrace{E(Y_{1i}|D_i=1)}^{\text{该去医院实际上也去了的人的身体状况}} - \overbrace{E(Y_{0i}|D_i=1)}^{\text{应该去医院但是没去的身体状况}}}_{\text{真实的因果关系（ATE）}} + \underbrace{\overbrace{E(Y_{0i}|D_i=1)}^{\text{应该去医院但是没去的身体状况}} - \overbrace{E(Y_{0i}|D_i=0)}^{\text{没去医院的人的身体状况}}}_{\text{样本选择偏误}} \tag{2}$$

什么是内生性问题

- 很大部分实证研究的目标就是为了克服样本选择偏误。以此获得：
  - 除了我们感兴趣的变量发生变化
  - 其他所有因素都一样
  - 并因此观测结果的变化
  - 从而识别出因果关系
- 方法：随机实验；控制实验；回归分析，等等。

什么是内生性问题

- 随机实验（医院的效果）

$$
\begin{aligned}
&E(Y_i|D_i=1)-E(Y_i|D_i=0)=E(Y_{1i}|D_i=1)-E(Y_{0i}|D_i=0)\\
&=E(Y_{1i}|D_i=1)-E(Y_{0i}|D_i=1)\\
&=E(Y_{1i}-Y_{0i}|D_i=1)\\
&=E(Y_{1i}-Y_{0i})
\end{aligned}
\tag{3}
$$

什么是内生性问题

- 随机实验的局限性：昂贵、周期长且操作困难。例子：Tennessee STAR experiment
  - 研究目的：估计班级大小对学生成绩的作用
  - 费用：1200 元美元
  - 研究对象：1985/86 学年入学的 11600 个小学生
  - 时间跨度：跟踪 4 年

什么是内生性问题

- 用回归分析的方法来制造“反事实”，这就回到了我们前几章节学习的内容了。

- Treat 与否一定独立于观测对象的自身属性 X
- 需要很多假设（固定线性模型的假设），可能与现实脱节
  - OLS 的一致性（伍德里奇，第五版，p138）

$$\hat{\beta}_1=[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i]\Big/[\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2]$$
$$=\beta_1+[n^{-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)\mu_i]\Big/[n^{-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2]$$

$$y_i=\beta_0+\beta_1 x_{i1}+\mu_i$$

$$p\lim\hat{\beta}_1=\beta_1+Cov(x_1,\mu)/Var(x_1)=\beta_1(Cov(x_1,\mu)=0)$$

  - 零均值和零相关

$$E(\mu)=0, Cov(x_j,\mu)=0$$

什么是内生性问题

$$Y_i = \begin{cases} Y_{1i} & D_i = 1 \\ Y_{0i} & D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i \tag{4}$$

$$Y_i = \underbrace{\alpha}_{E(Y_{0i})} + \underbrace{\beta}_{(Y_{1i}-Y_{0i})} D_i + \underbrace{\mu_i}_{Y_{0i}-E(Y_{0i})} \tag{5}$$

$$E(Y_i|D_i = 1) = \alpha + \beta + E(\mu_i|D_i = 1) \tag{6}$$

$$E(Y_i|D_i = 0) = \alpha + E(\mu_i|D_i = 0) \tag{7}$$

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \underbrace{\beta}_{TreatmentEffect} + \underbrace{E(\mu_i|D_i = 1) - E(\mu_i|D_i = 0)}_{SelectionBias} \tag{8}$$

$$E(\mu_i|D_i = 1) - E(\mu_i|D_i = 0) = E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0) \tag{9}$$

什么是内生性问题

- 对于给定的线性回归模型

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e \tag{10}$$

- 如果干扰项和解释变量是相关的，即

$$E(e|X_1, X_2, \cdots, X_k) \neq 0 \tag{11}$$

那么可以说这个线性模型存在内生性问题

- **内生性问题指当干扰项和解释变量相关时，我们无法识别解释变量的因果关系系数。**
- 内生性问题会造成最小二乘法估计系数有偏。
- 在因果关系分析中，偏差指相关关系系数对因果关系系数的偏差。
- 正如我们之前的例子中，我们期望估计出 β_1, 但实际是对 γ_1 的估计，而 $\gamma_1 = \beta_1 + \beta_2 \phi_1$。

内生性的来源

- 目前，内生性产生的原因主要有五个方面：

1. 遗漏变量偏误
2. 测量误差
3. 双向因果
4. 样本选择偏误
5. 函数形式误设

内生性的来源

- **遗漏变量偏误**。如能力，家庭背景...
  - 假设模型为：$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + e, E(e|X_1, X_2) = 0$，但回归估计的模型遗漏了 X_2，使用以下模型：$Y = \alpha + \beta_1 X_1 + \nu$。
  - 如果 X_1 和 X_2 存在相关性，那么：
    $E(\nu|X_1) = E(\beta_2 X_2 + e|X_1) = \beta_2 E(X_2|X_1) \neq 0$
  - 由此可见，遗漏变量造成内生性的原因是：遗漏变量和解释变量的相关性导致包含遗漏变量的干扰项 ν 和解释变量 X_1 相关。

内生性的来源

$Y_i = \beta_1 + \beta_2 X_{2i} + \underbrace{\beta_3 X_{3i} + u_i}_{u_i^*}$, which satisfies ideal conditions

X_{3i} is not observed or hard to measure

Run OLS of Y_i on X_{2i} with X_{3i} omitted, we have

$$\hat{\beta}_{2,ols} = \frac{\Sigma(X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\Sigma(X_{2i} - \bar{X}_2)^2}, \text{ and } Y_i - \bar{Y} = \beta_2(X_{2i} - \bar{X}_2) + \beta_3(X_{3i} - \bar{X}_3) + u_i - \bar{u}$$

Then,

$$\hat{\beta}_{2,ols} = \beta_2 + \frac{\beta_3 \Sigma(X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)/N}{\Sigma(X_{2i} - \bar{X}_2)^2/N} + \frac{\Sigma(X_{2i} - \bar{X}_2)(u_i - \bar{u})/N}{\Sigma(X_{2i} - \bar{X}_2)^2/N}$$

$$= \beta_2 + \beta_3 \frac{sample\,\mathrm{cov}(X_{2i}, X_{3i})}{sample\,\mathrm{var}(X_{2i})} + \frac{sample\,\mathrm{cov}(X_{2i}, u_i)}{sample\,\mathrm{var}(X_{2i})}$$

As $N \longrightarrow \infty$

$\hat{\beta}_{2,ols}$ tends to $\beta_2 + \beta_3 \frac{\mathrm{cov}(X_{2i}, X_{3i})}{\mathrm{var}(X_{2i})} + \frac{\mathrm{cov}(X_{2i}, u_i)}{\mathrm{var}(X_{2i})}$

$\hookrightarrow 0$

$p\lim \hat{\beta}_{2,ols} = \beta_2 + \beta_3 \cdot b_{32}$, b_{32} is *plim* of slope coefficient from OLS regression of X_{3i} on X_{2i}.

Generally, $\hat{\beta}_{2,ols} \neq \beta_2$, unless $\beta_3 = 0$ or $b_{32} = 0$, the latter one means $\mathrm{cov}(X_{2i}, u_i^*) = 0$

内生性的来源

- 没有足够控制变量时遗漏变量偏误的解决办法

1. 面板数据模型。面板数据可以控制不可观测的遗漏变量，但是要求这些遗漏变量不随时间变化。
2. 工具变量法 (IV)。
3. 随机控制实验（DID 或 RD）。
4. 时间差分和空间差分法（Duranton et al.,2009; Belotti et al.,2016）; 时间趋势多项式法（Wolfers,2006）; 交互固定效应（Bai,2009; Kim and Oka,2014）。

内生性的来源

- **测量误差**: 包含解释变量的测量误差和被解释变量的测量误差。如，受访者提供错误答案、数据录入错误等。
- **（1）解释变量的测量误差**

- 假设模型为：$Y^* = \beta_0 + \beta_1 X^* + e, E(e|X^*) = 0$。在观测 X^* 时存在测量误差 μ, 即观测到的 $X = X^* + \mu$, 但测量误差 μ 与 X^*、Y^* 不相关，且均值为 0。Y^* 观测值不存在测量误差，即 $Y^* = Y$。
- 把观测值代入模型：

$$Y^* = \beta_0 + \beta_1(X - \mu) + e = \beta_0 + \beta_1 X - \beta_1\mu + e = \beta_0 + \beta_1 X + \underbrace{\nu}_{-\beta_1\mu + e}$$

- 此时观测值 $X = X^* + \mu$ 和干扰项 $\nu = -\beta_1\mu + e$ 中都包含测量误差 μ。
- 这导致：

$$Cov(X, \nu) = Cov(X^* + \mu, -\beta_1\mu + e) = -\beta_1 Var(\mu) \neq 0$$

内生性的来源

- 由此可见，解释变量的观测误差造成内生性的原因是：使用有测量误差的解释变量造成了干扰项里包含测量误差，进而导致干扰项与观测的解释变量相关（因为二者都包含了测量误差）。
- 当样本容量很大时，该偏误依然存在。

- **（2）被解释变量的测量误差**

- 假设模型为：$Y^* = \beta_0 + \beta_1 X^* + e, E(e|X^*) = 0$。在观测 Y^* 时存在测量误差 μ, 即观测到的 $Y = Y^* + \mu$, 但测量误差 μ 与 Y^*、X^* 不相关，且均值为 0。X^* 观测值不存在测量误差，即 $X^* = X$。:
- 把观测值代入模型：$Y^* = Y - \mu = \beta_0 + \beta_1 X^* + e$

$$Y = \beta_0 + \beta_1 X^* + e + \mu = \beta_0 + \beta_1 X^* + \underbrace{\nu}_{\mu+e}$$

内生性的来源

- 这时干扰项 $\nu = \mu + e$ 中包含测量误差 μ，但 X^* 不存在误差。

$$E(\nu|X^*) = E(e + \mu|X^*) = 0$$

- 因此，二者不相关，即不存在内生性问题。不过由于误差项变大（方差变大），回归结果的显著性会有所降低。

内生性的来源

- 变量的测量偏误的解决方法
  - 工具变量法。工具变量与 X 相关，但不与测量误差相关。
  - 建立测量误差的数学模型，并且如果有可能，用得到的公式调整估计值。
  - 最简单的方法就是利用一个更精确的 X 来重新回归，也就是用多个表示 X 含义的变量来回归，并将结果进行比较。

内生性的来源

内生性的来源

- **双向因果**：若 X 导致 Y，Y 也导致 X。如果存在双向因果关系，则 OLS 回归中同时包含了这两个效应，此时 OLS 估计量是有偏、非一致的。

内生性的来源

- 考虑如下情形，Y_1 和 Y_2 互因果：

$$Y_1 = \beta_1 X_1 + \phi_1 Y_2 + e_1 \tag{12}$$

$$Y_2 = \beta_2 X_2 + \phi_2 Y_1 + e_2 \tag{13}$$

其中，X_1 和 X_2 均与 e_1 和 e_2 不相关，且 e_1 和 e_2 也不相关。

- 将方程（13）代入到方程（12）中，可以得到：

$$Y_1 = \frac{\beta_1}{1-\phi_1\phi_2} X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2} X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2} \tag{14}$$

内生性的来源

- 由上式可以看到：

$$
\begin{aligned}
Cov(Y_1, e_2) &= Cov(\frac{\beta_1}{1-\phi_1\phi_2}X_1 + \frac{\beta_2\phi_1}{1-\phi_1\phi_2}X_2 + \frac{e_1}{1-\phi_1\phi_2} + \frac{e_2\phi_1}{1-\phi_1\phi_2}, e_2) \\
&= \frac{\phi_1}{1-\phi_1\phi_2}Var(e_2) \neq 0
\end{aligned}
\tag{15}
$$

- 所以 Y_1 具有内生性问题。同理，Y_2 也具有内生性问题。

- 解决办法：工具变量法和随机控制实验

内生性的来源

- **样本选择偏误：**当抽样过程影响数据的可得性且与因变量有关时就产生了样本选择偏误。样本选择偏误导致一个或多个回归变量与误差项相关，因此使 OLS 估计量有偏且非一致。

情形 1 数据缺失完全是随机的。即缺失的原因是随机的，与 Y 或 X 的取值无关，这只会导致样本规模变小，并不会导致估计偏误。

情形 2 缺失数据依赖于某个回归变量的取值时。这种情形也只会导致样本规模变小，不会引起偏误。例如，只采用了学生/教师比超过 20 的地区。（只是不能得到比值小于 20 地区有关结论）

情形 3 由样本选择过程与因变量 Y 的取值相关联。这样的选择过程会引起误差项和自变量的相关，由此产生 OLS 估计量的偏误（样本选择偏误）。

- 解决方法：在实践中，常用的方法是改变样本，对不同样本进行回归，并比较回归结果。

内生性的来源

- **函数形式的误设**。
  - 这种偏差其实也是一种遗漏变量偏误，其中的遗漏变量为反映回归函数中缺少的非线性部分的项。例如，若总体回归函数为抛物线形式，那么线性回归模型就遗漏了二次项变量。
  - 函数形式误设通常利用观测数据和回归函数的估计图来发现，并采用另一种不同的函数形式进行修正。例如，多项式回归函数、对数形式、交互项、线性概率模型等。

内生性的来源

- OLS **标准误非一致**导致了另一种内部有效性的威胁。即使 OLS 估计量一致且样本较大，但标准误非一致会使假设检验的水平不同于“要求”的显著水平，而且使得置信区间在“要求”置信水平下没有包含真值。
  - 引起标准误非一致主要有两个原因：异方差；序列相关。
  - 解决异方差的方法是利用异方差稳健标准误，并利用异方差稳健方差估计量来构建 F 统计量。
  - 解决序列自相关的方法就是利用其他标准误公式来校正这个问题。

“合格”的控制变量

- 当遗漏变量可观测时或者有足够的控制变量时遗漏变量偏误的解决办法。如果我们有遗漏变量的数据，那么把它们放入多元回归即可。
  - 那么，是否应该在回归中包含更多的变量？
  - 并非控制变量越多越好。增加一个变量有得也有失：
    1. 一方面，增加遗漏变量会减缓潜在遗漏变量偏误；
    2. 另一方面，加入一个不重要的变量会降低其他回归系数估计量的精确度。

“合格”的控制变量

- **不合格的控制变量：**就是那些可以作为实验结果的变量，它本身可作为被解释变量的变量
- **合格的控制变量：**当选定回归元后，其取值已经固定给出的变量。
- 不合格的控制变量带来的问题——选择偏误。
  - 考虑大学教育对收入的影响，同时人们还可以在白领和蓝领之间进行职业选择。
  - 职业选择和教育水平、收入都高度相关，它是不是遗漏变量？
  - 存在的问题：考虑到教育水平影响职业选择，即使教育水平随机分配，同一职业内不同教育水平下的工资差异不可相互比较。

“合格”的控制变量

- **大学教育与职业选择问题：**设 W_i 表示个体 i 是否为白领工人，收入水平为 Y_i。接受或不接受大学教育都带来收入水平和职业选择的两种不同潜在结果，分布记为 $\{Y_{1i}, Y_{0i}\}$ 和 $\{W_{1i}, W_{0i}\}$。

- 于是，收入水平和职业选择的实现有赖于是否大学毕业和潜在可能结果的相互作用：

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$

$$W_i = C_i W_{1i} + (1 - W_i) Y_{0i}$$

- 其中，$C_i = 1$ 表示大学毕业水平，$C_i = 0$ 为其他。假设 C_i 随机分配，它独立与所有的潜在结果：

$$E(Y_i|C_i = 1) - E(Y_i|C_i = 0) = E(Y_{1i} - Y_{0i})$$

$$E(W_i|C_i = 1) - E(W_i|C_i = 0) = E(W_{1i} - W_{0i})$$

- 在实际操作中，分别将 Y 和 W 关于 C 回归就可以得到平均因果效应。

“合格”的控制变量

- 给定白领职业（$W_i = 1$），考虑大学毕业生和非大学毕业生的收入差距。
  - 我们可以在包含了 W_i 的回归模型中计算这个收入差距，也可以在 $W_i = 1$ 的所有个体中对 Y_i 关于 C_i 进行回归。后一个方法得到的估计值就是 $W_i = 1$ 时 C_i 取值分别为 0 和 1 时显示出的平均收入上的差异：

$$
\begin{aligned}
& E(Y_i|W_i = 1, C_i = 1) - E(Y_i|W_i = 1, C_i = 0) \\
&= E(Y_{1i}|W_{1i} = 1, C_i = 1) - E(Y_{0i}|W_{0i} = 1, C_i = 0) \\
&= E(Y_{1i}|W_{1i} = 1) - E(Y_{0i}|W_{0i} = 1) \\
&= \underbrace{E(Y_{1i} - Y_{0i}|W_{1i} = 1)} + \underbrace{\{E(Y_{0i}|W_{1i} = 1) - E(Y_{0i}|W_{0i} = 1)\}}
\end{aligned}
$$

这里用到了 $\{Y_{1i}, W_{1i}, Y_{0i}, W_{0i}\}$ 的联合分布与 C_i 相互独立的假设。

“合格”的控制变量

- 这个等式指出了不合格控制变量带来的问题，相比较的不是同类事物。
- 换言之，给定考虑的个体都是白领工人，是否拥有大学学历造成的工资差异等于大学文凭对 $W_{1i} = 1$(获得大学学历后成为白领工人) 的那些人带来的因果效应加上一个选择偏误，这一选择偏误反映出的是大学学历会改变白领工人组成这一事实。
- 这个例子的选择偏误符号不确定，依赖于职业选择、是否上大学以及潜在收入水平之间的关系。（选择偏误很可能是负的，一般大学毕业生获得白领工作应该是比较合理的，但是那些没有大学学历也能获得白领工作的人是比较“特别的”。）
- 因此，我们最好还是用不由教育水平决定的那些变量作为控制变量。

“合格”的控制变量

- **使用代理变量做控制变量：**即纳入回归方程的变量可能部分的控制遗漏变量，但是它本身被我们感兴趣的变量影响。

$$Y_i = \alpha + \rho s_i + \gamma a_i + e_i$$

- 控制变量为表示“能力”的数值变量 a_i, 可视为八年级时学生的智力检测分数，这个分数可以度量学生的天生能力。假设所有人都要完成八年的教育。$E(s_i e_i) = E(a_i e_i) = 0$。
- a_i 是个人做出教育 s_i 决策之前做出的。所以它是一个好的控制变量。
- 不幸的是，关于 a_i 的数据不可得。但是，我们可以用另外一个指标来代理，这个指标是在接受完教育后得到的（比如在求职申请中用到的测试分数）。可以理解为后天能力，记为 a_{li}。
- 一般，在先天能力的基础上，接受过教育后个体提高了他们后天的能力。

$$a_{li} = \pi_0 + \pi_1 s_i + \pi_2 a_i$$

“合格”的控制变量

- 当仅用 s_i 对 Y_i 进行回归时，你会担心遗漏变量引起的偏误。但是由于数据的可得性，我们用了代理指标进行回归。

$$Y_i = (\alpha - \gamma\frac{\pi_0}{\pi_2}) + (\rho - \gamma\frac{\pi_1}{\pi_2})s_i + (\gamma + \frac{\gamma}{\pi_2})a_{li} + e_i$$

- 在这个例子中，γ、π_1、π_2 都是正的，因此只有当 π_1 变得很小时，估计出的结果才接近因果效应。
- 如果用 a_{li} 对 s_i 的结果结果为零，那么你可以相当自信地假设上式中的 π_1 等于零。

“合格”的控制变量

- 对被解释变量添加不合格的控制变量

- 如果想在教育回归的研究获得带有因果性的结果，就不要再回归方程中加入职业作为控制变量。在代理变量作为控制变量的问题中，即使加入代理控制变量仍然没能得到我们感兴趣的回归参数，但相比于没有加入该变量，得到的结果改进了。
- 当不存在控制变量时，关于 s_i 做回归产生的系数是 $\rho + \gamma\delta_{as}$，其中 δ_{as} 是对 a_i 关于 s_i 进行回归得到的系数。
- 当使用代理控制变量得到的系数比没有控制变量情况下更接近于 ρ。

“合格”的控制变量

- **挑选控制变量的基本准则：考虑控制变量被决定的时间，一般在我们感兴趣的变量产生前就被决定的变量都是好的控制变量。**
  - 但是有时我们面对的控制变量被决定的时间不确定或未知。在这种情况下，因果关系的准确考虑需要我们做出哪个变量先被决定的假设，或者去说明没有任何一个控制变量是由我们感兴趣的变量所影响的。当然，也有一些其他的权衡依据：

1. 识别出回归模型中感兴趣的关键系数。如在测试成绩回归中，关键系数是学生/教师比系数。
2. 问你自己“这个回归中最有可能的重要遗漏变量偏误来源于哪里？”这就需要我们的经济理论和专业知识了，并且应该在建立实际回归前就加以考虑。由此即得基准回归模型的设定形式。
3. 利用第二步中确定的其他可疑控制变量扩展基准回归。如果加入的控制变量系数是统计显著的，或者加入遗漏变量后，感兴趣系数的估计量发生明显变化（估计量方差会增大），那么这些变量应该保留，反之亦然。
4. 用表格形式正确概括结果。这可以“揭示”读者的潜在疑虑。

参考文献：

Stock and Watson(2011), Introduction to Econometrics, 3rd edition.

Angrist and Pischke(2008). Mostly harmless econometrics: An empiricist’s companion. Princeton university press.

连玉君，Stata 讲义或者视频资料.

NBER 计量课程：*http : //www.nber.org/SI_econometrics_lectures.html*