第五讲 回归分析评价框架

许文立*[1,2]

[1]安徽大学经济学院
[2]安徽生态与经济发展研究中心

January 6, 2018

回归分析已经成为计量经济学领域最重要的经验研究方法。那么，自然出现的问题就是：我们如何评价基于回归分析的经验分析呢？或者如何判断采用回归分析方法的经验研究是否可信呢？

从第三讲和第四讲可以看出，一元回归会遗漏重要的回归量，从而导致我们关心的效应产生遗漏变量偏误，引入数据可用的遗漏变量，采用多元回归可以消除遗漏变量偏误。那么，如何评价我们所做的回归分析呢？

在本讲中，我会向大家介绍评价一个有用的经验研究的标准和步骤。如果发现回归分析的问题，应如何改进。

1 内部有效性和外部有效性框架

评价一个回归分析的有效性，要基于内部有效性和外部有效性的概念。如果一个研究关于因果效应的统计推断对于所研究的总体和环境是有效的，那么，这个研究就具有**内部有效性**；如果这个研究德尔结论能推广至其他总体和环境，那么，它也具有**外部有效性**。

其中，**研究的总体**是研究所刻画的样本来源总体；**感兴趣的总体**是从研究中得到的因果推断推广应用的总体。例如，高中各种实验班对211、985高校升学率的效应，是否可以推广至高校各类实验班的设立呢。

而“环境”则是指制度、法律、社会、文化和经济环境等。例如，前面回归所得的美帝小班教学的效应，是否对中国有用呢？因为美帝和中国差异还是非常大的。

1.1 内部有效性

内部有效性由两部分构成：

1、因果效应估计量是无偏和一致的。

2、假设检验有合意的显著性水平，并且置信区间有合意的置信水平。

*E-mail：xuweny87@163.com。

在回归分析中，因果效应是利用估计的回归函数来估算的，假设检验是利用估计的回归系数和标准误来执行的。因此，内部有效性要求OLS估计量是无偏和一致的，标准误的计算要使得置信区间有合意的置信水平。但是实践中，有许多原因使得这个要求不能得到满足。我们第四讲中提到的遗漏变量偏误就是其中之一，因为它使得回归量与误差项相关，破坏了OLS第一假设。如果遗漏变量数据可用，我们可以纳入回归模型中来消除遗漏变量偏误。其它原因，我们在下面将会详细讲解。

1.2 外部有效性

从外部有效性的定义可知，破坏外部有效性的潜在因素来源于所研究总体和环境与感兴趣的总体和环境存在差异。

1、总体差异

所研究的总体与感兴趣的总体之间存在差异会威胁到一个研究的外部有效性。例如，医药实验总是从小白鼠开始，但是一种新药在小白鼠身上起作用，其对人类也起作用吗？毕竟小白鼠总体与人类总体存在非常大的差异。

一般来说，真实因果效应在所研究的总体和感兴趣的总体中不可能完全相同。

2、环境差异

即使所研究的总体和感兴趣的总体相同，只要环境存在差异，一个研究的结果也不可能推广到更一般化情形。

3、加利福利亚的小班教学

从三、四讲来看，加利福利亚州的小班教学确实会提高学生的平均测试成绩，也就是说，学生-老师比越小，平均测试成绩越高。但是，这一结果是在加利福利亚的初等教育学生这一总体得出的。如果想推广至中等教育，甚至高等教育，小班教学的回归分析可能就存在外部性有效性问题，因为初等教育总体和中等教育、高等教育总体存在差异。

同样的道理，如果想把这一研究结果应用到中国来，也可能存在外部有效性问题。

综上所述，所研究的总体和环境与感兴趣的总体和环境越接近，外部有效性就越强。

4、如何判断外部有效性？

要判断外部有效性，可能就要我们非常了解所研究总体和环境与感兴趣的总体和环境。两者之间的重要差异都可引起对研究外部有效性的怀疑。

从实践的角度来说，如果我们有不同但相关总体的多个研究，那么，外部有效性就能通过对比这些研究结果来判断。一般来说，多个研究得到相似的结果可以支持外部有效性，反之亦然。

5、如何设计一个外部有效的研究

因为外部有效性来源于总体和环境之间的不可比较或者比较起来较为困难。因此，最小化外部有效性的威胁要在研究初期解决，例如，研究设计和数据收集阶段。有关研究设计的讨论可以参见Shadish，Cook，Campbell（2002）。

2 内部有效性的威胁

因为回归分析的内部有效性包含两个方面的内容：OLS估计量无偏和一致；标

准误得到的置信区间有合意的置信水平。

目前，回归分析中引起估计量有偏的原因主要有五个方面：遗漏变量偏误、回归函数误设、变量的测量误差、样本选择偏误、双向因果。出现这五个方面的问题都是因为总体回归方程中回归量与误差项相关，打破了OLS第一假设。

2.1 遗漏变量偏误

回忆一下第四讲中提到的遗漏变量偏误。遗漏变量既要决定Y，又要与X相关。即使在大样本下，遗漏变量偏误也会存在，因此，OLS估计量不具有一致性。如何最好地最小化遗漏变量偏误取决于控制的潜在遗漏变量是否可用。

1、当变量可观测或有恰当控制的变量时，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，把它们包含进多元回归中即可。

但是在回归模型中，增加一个变量既有好处又有坏处：一方面，遗漏变量会导致遗漏变量偏误，增加遗漏变量会消除潜在的遗漏变量偏误；另一方面，包含一个不重要的变量（例如，它的总体回归系数为0）会降低另一些回归因子系数估计量的精确性。**换句话说，是否加入一个变量，就等同于在系数的偏误和方差之间做出取舍。**实践中，通常用以下四步哎判断是否要加入一个变量或一些变量：

第一步,识别出回归模型中感兴趣的系数和关键系数。例如，前面的例子中，关键系数是学生-老师比的系数。

第二步，自问“我们的回归中，重要的遗漏变量偏误来源于什么？”这就需要我们应用经济理论和专业知识。这一步应该在回归之前就完成。这一步的回归应该当做**基准回归模型**，也就是经验回归分析的开始。

第三步，扩展基准回归。如果加入的控制变量系数是统计显著的，或者如果加入遗漏变量后，感兴趣的变量系数估计量发生明显变化，那么，这些变量应该保留，反之亦然。

第四步，把我们的回归结果详细的列示在图表中。这一步是为了方便别人查看，并发现一些可疑之处，进而提出一些有益的意见和建议。

2、当适当的控制变量不可用时，遗漏变量偏误的解决办法

当遗漏变量的数据不可用时，我们就不能将其增加到回归模型中作为一个回归量。但是，实践中还有以下几种方式来解决遗漏变量偏误。每种方法都适用不同的数据类型。

第一种方法：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**。

第二种方法：工具变量法，这种方法依赖一个被称为**工具变量**的新变量。

第三种方法：随机控制实验——DID或者RD等。

在实践中，以上三种方法通常结合使用。但目前的主流面板数据模型仍只考虑了不随时间变化的不可观测遗漏变量。那么，如果这些遗漏变量随时间可变，即属于不可观测的时变特征，解决这类问题的方法如下：

第四种方法：时间差分和空间差分法（详见Duranton et al.，2009；Belotti et al.，2016）、时间趋势多项式法（包括时间趋势二次型）（参见Wolfers，2006）、交互固定效应（参见Bai，2009；Kim and Oka，2014）。

后面几讲中，我们将详细讲解前三种方法。

2.2 函数形式误设

我们前面将的回归函数是线性的，但是如果真实总体回归函数是非线性的，那么，这种函数形式误设也会使得OLS估计量有偏。为什么这种偏误也属于遗漏变量偏误的一种类型呢？试想一下，如果总体回归函数是抛物线，那么，线性回归模型就遗漏二次项变量，这个二次项变量也当做一个回归量，就与上面的遗漏变量偏误没有本质区别。

函数形式误设通常利用图形和估计的回归函数来甄别。它能利用不同的函数形式进行纠正。例如**多项式回归函数、对数形式、交互项、线性概率模型（probit或tobit 等）等等**。

2.3 测量误差

加假如我们在测量或者录入数据时，不小心把自变量的数据搞错了，者也会导致OLS估计产生偏误。这种情形称为**变量误差偏误**。

变量误差可能有许多原因：调查时，受访者提供错误答案；数据录入错误等等。从数学形式来讲，ΔX_i表示测量误差，那么，根据测量误差修正总体回归方程$Y_i = \beta_0 + \beta_1 X_i + u_i$

$$Y_i = \beta_0 + \beta_1(X_i + \Delta X_i) + u_i = \beta_0 + \beta_1 X_i + \beta_1 \Delta X_i + u_i \tag{1}$$

从式（1）中可以看出，存在变量误差时，就相当于多元回归中遗漏了误差这个自变量，从而引起遗漏变量偏误。

如果Y存在测量误差，那么，会使得回归方差增大，但是不会引起β的估计偏误。例如，Y的误差会进入随机误差项，即新的误差项$v_i = u_i + w_i$。如果w_i是随机的，那么，$E(w_i|X_i) = 0$，这也意味着$E(v_i|X_i) = 0$，因此，β的估计量是无偏的，但是$var(v_i) > var(u_i)$，使得β的估计量的方差增大。

解决方法

方法一：工具变量法，工具变量与X相关，但不与测量误差相关。

方法二：提出一个测量误差的数学模型，如果可能用包含这个数学公式所表示的指标来调整估计量。最简单的方法就是利用一个更精确的X来重新回归。也就是通常，用多个表示X 含义的变量来回归，并将结果进行比较。

2.4 缺失数据和样本选择

数据缺失是经济数据集的常见现象。数据缺失是否对内部有效性造成威胁取决于数据为什么缺失。考虑三种情形：

情形一：数据缺失完全是随机的。在这种情形下，随机缺失的原因与X或Y不相关，这只会导致样本规模变小（删除缺失样本），而不会导致估计偏误。

情形二：缺失数据依赖于X。在这种情形下，也只会导致样本规模变小，不会引起偏误。

情形三：由于样本选择过程造成的数据缺失，与Y相关。这种选择过程会引起误差项与回归量相关。这种OLS估计量偏误称为**样本选择偏误**。

解决方法：上面提到的那些方法都解决不了样本选择偏误。在实践中，常用的方法是改变样本，对多个样本进行回归，并比较回归结果。

2.5 双向因果

迄今为止，我们都假设因果关系只从X到Y，即X导致Y。但是，如果因果关系也从Y到X呢？也就是，X导Y，Y也可以导致X，这就是**双向因果关系**。双向因果也会导致误差项与回归量相关。用数学形式描述反向因果联系：

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{3}$$

公式（2）表示X变化引起的Y的效应，公式（3）则是反向因果。想想一下u_i 为正，那么，Y_i会增大。而反向因果关系表明，更高的Y会影响X的值。如果γ_1为正，那个Y越大，X也越大，那么，X_i与u_i 就正相关。$cov(X_i, u_i) = cov(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 cov(Y_i, u_i) + cov(v_i, u_i) = \frac{\gamma_1 \sigma_u^2}{1-\gamma_1 \beta_1}$

解决方法

有两种方法可以解决双向因果引起的偏误：**工具变量回归；随机控制实验**。我们将在后面详细讲述。

2.6 OLS标准误的不一致性

不一致的标准误会对内部有效性造成很大的伤害。即使OLS估计量是一致的，样本规模也很大，但是非一致的标准误会导致假设检验的显著性水平不合意，并使得置信区间在合意置信水平下没有包含真值。

引起非一致标准误，主要有两个原因：

1、异方差

有些统计软件只呈现同方差标准误。但是，如果存在异方差，标准误就不能作为可信的假设检验和置信区间的基础。

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。Stata软件提供了稳健标准误和稳健的F统计量。

2、序列相关

在某些情形下，总体回归误差在观测值之间是相关的。如果数据是完全随机抽样得到的，那么，序列相关就不会发生。但在实践中，经济样本往往是部分随机的。

序列相关并不会使得OLS估计量产生偏误或者非一致性，但是它会打破OLS第二个假设。这就会使得估计的OLS标准误产生的置信区间不在合意的置信水平下。

解决方法就是利用另一种标准误的计算公式。这些将在面板数据模型中详细阐述。

3 小班教学及其Stata操作

我们将上述外部有效性和内部有效性框架应用于小班教学效应研究。

3.1 外部有效性

前面两讲中使用的美帝加利福利亚420个学区的缩减班级规模对测试分数的影响是否能一般化到美帝其它州呢——即是说，这项研究是否具有外部有效性？那么，我们就要看看加利福利亚的学区总体及其环境是否能推广至其他州。

1.2节给出了一种判断外部有效性的方法：比较两个或多个相同主题研究的结果。因此，我们再利用另一个州——马萨诸塞州——的初等教育标准测试得分数据的回归结果来与加利福利亚州的回归结果进行对比。

数据比较。首先，比较变量的定义，即两项研究主要的变量定义比较接近。其次，比较两个样本的主要统计量，如表1所示。

Table 1: 样本比较

	加利福利亚		马萨诸塞	
	均值	标准差	均值	标准差
测试分数	654.16	19.05	709.83	15.13
学生-老师比	19.64	1.89	17.34	2.28
非英语母语学生比(%)	15.77	18.29	1.12	2.90
免费午餐学生比(%)	44.71	27.12	15.32	15.06
地区平均收入(千元)	15.32	7.23	18.75	5.81
样本量	420		220	
年份	1999		1998	

从表1看出，加州的平均测试分数更低，但是由于两个州测试不同，这种直接比较没有多大意义。加州的平均学生-老师比也更大，也就是说，加州的平均班级规模更大。加州的平均收入更低，但是标准差更大。而加州的非英语母语学生比例以及免费午餐学生比例均更高。

4 总结

1、内部有效性与外部有效性框架

内部有效性根据因果效应的统计推断来评价；外部有效性根据是否具有一般性来评价。因此，外部有效性要在研究设计或数据收集之前完成，而经验研究中更多关注于内部有效性评价。

2、内部有效性评价

内部有效性的威胁及其解决措施：

（1）遗漏变量偏误：（a）对于可观测的变量，加入可能减低偏误，但是估计量的方差会增加，四步走：第一步，识别你的核心系数或感兴趣的系数；第二步，根据经济理论和专业知识、经验，自问重要的遗漏偏误来源可能有哪些；第三步，检验第二步中那些仍有疑问的控制变量系数是否为0；第四步，把所有回归结果都呈现出来，让同行给你意见或建议。（b）对于不可观测的遗漏变量，三种解决办法：第一，面板数据；第二，工具变量；第三，随机控制实验（DID、RD等）。

（2）回归函数误设

有专门的非线性回归解决方法

（3）测量误差

最好的办法就是得到更精确的变量值。但这通常是不可能的，那么，两种方法备选：第一，工具变量，与变量相关，但与误差无关；第二，构建测量误差的数学公式来调整估计值，也就是说换一个变量值。

（4）样本选择偏误

（5）反向因果或同时因果

两种方式解决：第一，工具变量；第二，随机控制实验。

（6）估计量不一致性来源

异方差：用异方差稳健标准误和异方差稳健方差估计量构建的F统计量来解决；

系列相关：使用稳健标准误来解决。