高斯马尔可夫条件估计量是blue（无偏）

从最根本的定义来说，内生外生首先是取决于系统的，在一个系统内部决定的变量，自然就是内生变量，在系统之外决定的变量，就是外生变量。比如，给一个系统，比如地球，那么当前情况下地球上一切可以统计的变量都是内生变量，但是阳光就是外生变量。那么如果以太阳系为研究的系统，那么自然，阳光此时也是内生变量了。  
这样说内生性外生性似乎很容易理解，但是涉及到经济问题似乎不是那么好办了，因为经济系统中，所有的变量很难说是完全独立的，比如货币发行量，似乎是央行决定，按理说是外生的吧，但是慢着，央行的货币不是随便发的，也是因为有经济体有需求才会向社会发行货币，这个就是货币外生和货币内生的讨论，研究的文章有很多。  
还是先说外生性吧，Leamer定义，如果y对x的条件分布（这个就是给出x值，对应随机变量y）不随x的生成过程的修正而发生变化，那么x就是外生变量。外生性似乎还是可以分为两类，前定性（前定变量是指独立于方程中同期和未来误差项的变量），严格外生（严格外生变量是指独立于方程中所有同期、未来， 和过去误差项的变量）。  
依照这个定义，我什么也看不出来，倒是可以从CLRM假定cov(Ut,Xt)≠0情况考虑。既然cov(Ut,Xt)≠0可以叫成内生性，那么 cov(Ut,Xt)=0大概可以叫外生变量了吧。chris的书前面把这个假定强化为X是非随机变量，当显然这一假定是靠不住的，X更多情况下是随机变量。这里涉及到前面曾经困惑的一个概率，随机解释变量，随机解释变量就是说解释变量是随机的，原因根据我的思考总结，大概是这两类，1.观测值存在误差2.根据Y=α+θX+μ，如果Y能影响X，由于Y是随机的，自然X也就带有随机性了。  
随机解释变量容易带来内生性的问题，但却也不是必然，比如X是随机解释变量，但是X和u是独立的，也就是说cov(Ut,Xt)=0的时候，是不违背 CLRM假设的。其实到这里，我们讨论的一切，什么内生性，自相关，异方差，这些为什么要讨论呢，就是因为我们经常用OLS模型进行估计，而CLRM的五个假定就是为了使得OLS的估计具有一致性，无偏性，有效性。这时候，你看，即使X是随机变量，如果 cov(Ut,Xt)=0，那么是用OLS模型估计的值仍然是具有上面三条性质的，也就是说回归没有问题。什么时候会出问题，cov(Ut,Xt)≠0， 这个时候的回归就不是一致的了，这个可以从无偏定义推出来通过估计式，然后将cov(Ut,Xt)≠0代入入就可以发现E(β)≠β等等类似，这个不是今天论述的重点。  
顺带第一下CLRM第五条假设，残差u服从正太分布，这个和估计值的一致无偏有效性没有关系，但是在进行有效性检测是，就要用到这个假设了，如果残差不符 合正太分布，根本就没办法进行任何检测了。不过好在根据中心极限定理，样本够多的时候，可以渐进趋向正太分布，拿来做有效性检测也不会有太大问题。  
回到内生性的话题，解释变量的内生性指的是模型中的解释变量与扰动项相关，这个问题可以的原因大体有以下几条：  
1 模型设定偏差，遗漏了变量，这样，被遗漏的变量就被放进了残差项了，如果对被解释的变量和其他解释变量相关，自然，就会出现cov(Ut,Xt)≠0也就是内生性问题了。  
2测量误差，测量误差也有两种，一种是对被解释变量Y的测量误差，这个其实不会引起内生性，另一种是对解释变量X的测量误差。说明一下，被解释变量Y的测 量误差，设y的真实值y\*,测量值y,测量误差e0=y-y\*,假设理论的回归方程为y\*=β0+β1x1+….将测量误差带入方程得 y\*=β0+β1x1+….ε+e0=y\*=β0+β1x1+….ν其中ν=ε+e0表示实际回归方程的残差，显然由于y的测量误差和xi是相互独立的， 那么实际回归方程的残差v也与各解释变量相互独立（无关），所以还是满足外生性的。至于解释变量x的测量误差，回归式y=β0+β1x1+….ε中，测量 误差产生于xk, ek=xk-xk\*，将测量误差带入回归式y=β0+β1x1+….ε+βke那么如果cov(x\*k,ek)=0，那么cov(xk,ek)=cov(x\*k+ek,ek)=σ^2,此时的测量误差便会引起内生性的问题了。  
3双向交互影响（或者同时受其他变量的影响）这种情况引起的内生性问题在现实中最为常见。其基本的原理可以阐述为，被解释变量y和解释变量x之间存在一个交互影响的过程。x的数值大小会引起y取值的变换，但同时y的变换又会反过来对x构成影响。这样，在如下的回归方程中：y=β0+β1x1+…+ε。如果残差项ε的冲击影响了y的取值，而这样的影响会通过y传导到x上，从而造成了x和残差项ε的相关。也就是引起了内生性问题。这里举几个简单、但经常遇到的例子说明。例1：金融发展与经济增长；例2：外商直接投资FDI与经济增长；例3：犯罪率与警备投入。而我们通常最难以确定的内生性问题就是这个问题，因为经济学领域，变量大多都是相互影响的，毕竟都是在这样一个社会系统里面。而我们之前那些讨论的内容，通常都不是我们遇到的主要问题，因为测量误差是既成事实，改进很困难，而遗漏变量，这个是模型设定的时候，如果你遗漏了，别人挑刺也很难找出你遗漏了什么。倒是最后的交互影响，因为几个解释变量放在那里，他想挑刺就说内生性，因为经济系统内部内生性很多，如果你不能很好的解释、检验，这一道关就不能让别人信服。  
 网上有人说，内生变量和外生变量很好区别：外生变量就像函数中的参量一样，不受模型内部变量的影响。而内生变量受模型中的变量影响，比如 X+Y=a  （1）X-Y=1    （2）将上面两个方程联立，可以得到一个模型，其中变量X 和Y是可以通过解方程从“模型中”解出的，故是内生变量。但是变量a无法通过解方程来决定，只能由外在因素决定，所以是外生的，相当于一个参数。这种说法倒是简单明了，很容易理解，也符合我们的直觉，但是在简历模型的时候，很多时候没有办法就是说a他就是外生的，他不是另外一个变量Z。  
处理内生性的问题通常用工具变量法，或者两阶段最小二乘，这个和对内生性概念的理解就没什么关系，就不提了。

解决内生性方法大致有以下几种：  
1.引入工具变量，用2SLS 或3SLS解决。  
2.因变量滞后一期，或者是因变量采用多年的数据的均值。  
3.因变量采用行业调整的均值。

Ols估计中的几个经典假设

这个我认为应该从导致内生性的原因来考虑，导致内生性主要有两个原因：  
1：遗漏变量，且遗漏变量与引入模型的其他变量相关。  
2：解释变量和被解释变量相互作用，相互影响，互为因果。  
如果是1导致的内生性，则工具变量肯定要与误差项无关了，不然回归系数的估计值是不一致的；  
如果是2导致的内生性，则可以理解为工具变量要与因变量无关.

工具变量就是在回归方程中加入一个与应变量无关而和自变量有关的变量来解决内生性问题。  
“工具变量的选择只要掌握一个关键点就行：找一个和内生性变量有数据相关的，但是和残差没有关系的东西，这就是你的IV了。例如贸易量如果是内生的，那么你找地理距离作为IV。北京到纽约的距离，那是自然形成的，没人认为是由你的Y或者残差导致的。但是你会发现贸易量和地理距离在数据上具有相关性。这就很好。这种数据相关性越强，IV的效果就越好。就这么一段话，IV变量回归就讲完了。在STATA里面，你直接把原回归方程写出来，然后把IV填进去就可以了，回车就得到你的结果。关键是你不一定能找到这样的工具变量。你能找到，这个工具也不大能用。不过要注意，IV不灵不代表你不能发表。你只要找到一个IV，效果不是差的太离谱，一般都能发。当然不能发国际一流了。国内是没问题。

某一个变量与模型中内生解释变量高度相关，但却不与随机误差项相关，那么就可以用此变量与模型中相应回归系数的一个一致估计量，这个变量就称为工具变量，这种估计方法就叫工具变量法(IV Method)。  
作为工具变量，必须满足下述四个条件：  
　　（1）与所替的内生解释变量高度相关；  
　　（2）与随机误差项不相关；  
　　（3）与模型中其他解释变量不相关；  
　　（4）同一模型中需要引入多个工具变量时，这些工具变量之间不相关。  
举个例子：比如，令GDP增长率为[被解释变量](https://www.baidu.com/s?wd=%E8%A2%AB%E8%A7%A3%E9%87%8A%E5%8F%98%E9%87%8F&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dWP1RsnhwBPAndPyn4n1wb0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPjnzPjR3n1b4" \t "_blank)，需要研究GDP增长率与出口开放程度的关系，可以引入工具变量“各省区到海岸线的距离”来替代“出口开放程度”。认为：1.各省区到海岸线的距离与各省区的出口密切相关；2.各省区到海岸线的距离与随机误差项无关。

1.混合估计模型就是各个截面估计方程的截距和斜率项都一样，也就是说回归方程估计结果在截距项和斜率项上是一样的。

1. 随机效应模型和固定效应模型则认为回归方程估计结果在截距项和斜率项上是不一样的，所以你可以选择变截距模型，也可以选择变系数模型
2. 随机效应和固定效应模型的区别在于，随机效应模型认为误差项和解释变量不相关，而固定效应模型认为误差项和解释变量是相关的

短面板好像一般很少用随机效用模型的。我很少看到文章做hausman检验来确定到底该用哪种模型，而且大部分都是用的固定效用模型。

我想可能的原因是：一方面，固定效应模型可以将扰动项中那些，个体异质和时间异质的，同时有可能导致内生性问题的误差估计出来，提高估计结果的一致性。而随机效应模型仅仅只是在无内生性的条件下，得到一个比固定效应更有效的估计量。

另一方面，本身huasman检验中，接受随机效应模型假设的前提是，随机效应模型和固定效应模型的估计量都是一致估计量。我想一致性的重要程度要高于有效性。如果估计量都估不准，再有效又有什么用呢？ 所以，个人觉得固定效应模型是个好东西！

固定效应模型是指实验结果只想比较每一自变项之特定类目或类别间的差异及其与其他自变项之特定类目或类别间[交互作用](http://baike.baidu.com/view/635317.htm" \t "_blank)效果，而不想依此推论到同一自变项未包含在内的其他类目或类别的实验设计。

随机效应模型(random effects models)是经典的线性模型的一种推广，就是把原来（固定）的回归系数看作是随机变量，一般都是假设是来自正态分布。

固定效应模型认为误差项和解释变量是相关的，而随机效应模型认为误差项和解释变量不相关。

固定效应和随机效应是在面板模型中关于扰动项的不同设定。二者都假定扰动项由两部分组成，一部分随时间也随个体而变，另一部分只随个体而变不随时间而变，所以反映了个体效应。  
随机效应说，这个个体差异部分的扰动项和解释变量不相关（外生），因此直接ols估计也可以得到一致估计，但是可以证明，这时候协方差矩阵不满足球形扰动，所以这样的估计不是有效率的，需要通过gls估计。  
固定效应说，这个个体差异部分的扰动项和解释变量相关（内生性），这就是典型的内生性问题，这时候按照上述方法直接估计的结果是不一致的。通过组内离差消除掉这个不随时间而变的麻烦的东西，就可以得到一致估计了。  
总结一下，当个体效应是外生的时候，固定效应和随机效应估计都是一致的，但随机效应更有效；当个体效应是内生的时候，随机效应的估计是不一致的，但固定效应是一致的。所以固定效应比较稳健。（选择固定效应与随机效应的关键键就是如何确定个体效应是否内生）  
我们可以根据“个体效应是外生的时候，固定效应和随机效应估计都是一致的”来构建统计量进行检验到底个体效应是不是外生的，因此如果是，那么这两个估计的差应该很小。这就是典型的检验内生性的方法，豪斯曼检验。

内生性问题影响因果判断

当前主流定量社会科学研究领域中，因果关系的判定主要基于反事实框架。即一个影响因子或者干预对个体的因果效应，应是该个体在控制组和干预组中两个可能的结果状态之间的差异。如同“人不能两次踏进同一条河流”一样，个体的结果只能在一个组中被观察到。这样，对于该个体而言，干预结果与控制结果中必有其一是缺失数据。谢宇曾以大学教育为例说明该问题：在分析大学教育对于个体的收入是否有因果效应时，对一个上大学的学生，我们不可能获得他不上大学的情况的数据。因此，社会学定量分析中只能用平均干预效应来替代，估算一组大学生（干预组）与一组非大学生（控制组）之间的平均收入差异。但替代的前提是，干预组和控制组必须在其他收入因素上是一致的，即两组人的年龄、性别、家庭背景、智商、性格等等的平均值全部相同。一旦两组之间在某个变量“E”的均值上不一致（例如性格，往往难以采集数据），且这个遗漏掉的变量本身和解释变量又有关系，那么这样估算出来的就是有偏差的，甚至是伪相关，因果判断就无从谈起。

回归分析一般通过控制一系列变量，使两个组别具有可比性。但是，总有些无法被观察或学者没有想到的变量，导致两组之间不具有可比性。对回归方程而言，这意味着解释变量和遗漏误差项出现相关，不能满足高斯马尔可夫定理，估计参数出现偏误。在计量经济学中，该问题被称为内生性问题。由于绝大多数实证研究都是基于非实验性数据，无法保证实验组和控制组的相似性，因此所有基于调查数据的实证研究，无一例外都会受到内生性问题的困扰。而这主要源自一般性的遗漏变量偏误（如无法测量的个人能力往往被方程遗漏）、自选择偏误（如分析高等教育的因果作用，但是否接受高等教育本身是个人选择）、样本选择偏误（如分析找熟人对于求职的因果作用，但找熟人求职的这类人本身可能具有某种特性）和联立性偏误（自变量和因变量之间有双向关系）等多方面因素的影响。

  既有研究对内生性问题关注不足

  社会学的研究目的在于弄清某种机制或得出因果关系，而非局限于描述自变量和因变量之间的统计相关，所以，内生性偏误就应该成为社会学定量研究中必须直面的问题。尽管既有研究对反事实因果关系、模型识别策略的关注一直存在，但在大量的实证研究中，内生性问题往往被一笔带过，甚至不加提及。国内 学者王天夫和章奇等人对社会科学领域因果分析的基本概念进行了细致的回顾，但对和因果推断紧密相关的内生性问题及其解决方案却没有进行深入探讨。

社会学定量研究中，内生性问题的解决方法可以概括为四个方面。第一，解决遗漏变量偏误。主要包括四种策略：一是利用非传统数据作为遗漏变量的替代，以控制潜在的遗漏变量；二是依赖于变量在时间维度上的差异，运用固定效应模型控制时间固定的非观测因素；三是利用组内策略，即用非时间性的组内差 异来估计固定效应模型；四是使用工具变量。

第二，解决自选择偏误（是可以的）。最直观的方法是使选择行为不存在，因此随机分配是解决此问题的最佳途径，因为它可确保主解释变量和未观 察因子之间没有任何关联性。解决自选择偏误的方法主要有三种：一是实验或自然实验。在社会互动研究中，外生性干扰（如自然实验、政策介入或自然发生事件） 可以用来建构排除限定（即工具变量）以帮助识别模型。二种是匹配法，也即倾向评分方法（PSM）。其理念是，用尽量相似的控制组成员和对照组成员来进行比较以减少偏误。三是建立选择过程模型。

第三，解决样本选择偏误。被广泛运用来解决样本选择问题的方法是赫克曼（James Heckman）的两阶段法。该方法的主要特点是简明性。但该方法也存在二分因变量的误用、计算反向Mills比率的难度，错估标准误差，以及真实模型的修正项与其他回归量之间共线性等问题。

第四，解决联立性偏误。联立性偏误实际也可以看成是一种特殊的遗漏变量偏误，因此工具变量方法显然适用。近来研究常常采用外生性政策的干预或者对部分样本的外生干扰来充当工具变量。另外，最新的文献提出了“条件方差限定”的模型识别策略。其基本思想是：对一个既定的社会互动现象，个人结果的 跨组别方差由三个因素所导致，即组群异质性方差、个体异质性方差与组群解释变量方差。如果每个组群由大小不同的次级组群随机组成，那么显然对次级组群而 言，上级组群和个体的异质性方差是一样的，但它们各自的解释变量方差却不一样。这样，通过比较不同次级组群的解释变量方差，就可以识别解释变量的因果效 应。

提升解决内生性问题能力

在借鉴前人研究和其他学科方法的基础上，社会学研究加强对内生性问题的解决能力，主要有四种路径。

第一，提高数据质量。无论是固定效应方程、匹配方法，还是工具变量分析等策略，都离不开高质量和丰富的调查数据。以“社会网”研究为例，以 往个体的社会网数据收集仅有两层数据——被访者和被访者的朋友。在今后的调查中，除以上信息外，还应该设法对“朋友”的配偶、父母和朋友的朋友进行数据收 集，构成一个三层数据体系：被访者（第一层）、被访者的朋友（第二层）与被访者的朋友的其他社会关系（第三层）。而这在本质上就为“朋友”提供了一个工具 变量，即第三层作为第二层的工具变量。因为第三层变量显然和第二层有关，但如果能确保第三层“朋友”的社会关系和被访者不直接认识（比如，被访者会听说自 己的朋友的某些社会关系人的信息），那么第三层变量就会仅仅通过影响第二层变量来影响第一层变量。

第二，对外生性数据的收集和使用。兄弟姐妹的数量、子女的性别、彩票号码、河流数目、政府外在政策干预等都是外生的数据，能够对社会资本研 究中内生性问题的解决起到作用。特别是工具变量往往需要“灵感式”的逻辑推导，一些看似和本研究无关的变量，反而有可能成为解决内生性问题的奇兵。

第三，根据不同的数据条件，运用适当的模型。例如，如果获得面板数据，则能理想地解决时间固定效应问题，在很大程度上提高因果推断的可靠性。而较大的样本，则可以考虑匹配方法，因为它只有在大样本情况下才能获得可靠的推断。

第四，明确说明运用这些模型和识别方法所需要的假设。固定效应方程和倾向性评分匹配能够部分解决内生性问题；工具变量以及自然实验则可以在 理论上全部解决内生性问题。但是，工具变量很难寻找，其外生性无法证明，且必须详细讨论弱工具变量和不合格工具变量的可能性；而实验方法具有一定的局限 性。此外，如果未对内生性问题进行处理，则必须说明获得的统计参数只是一个描述性的统计关系，而不能断言是因果关系。