

1. 证明在高斯马尔科夫条件下的估计量是BLUE的。误差零均值，同方差，且互不相关的线性回归模型中

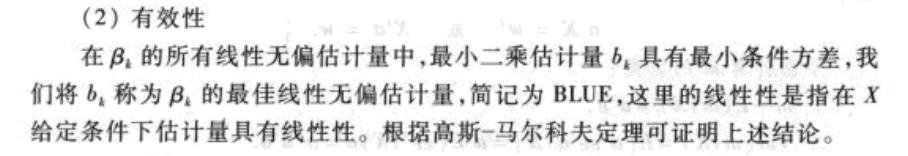
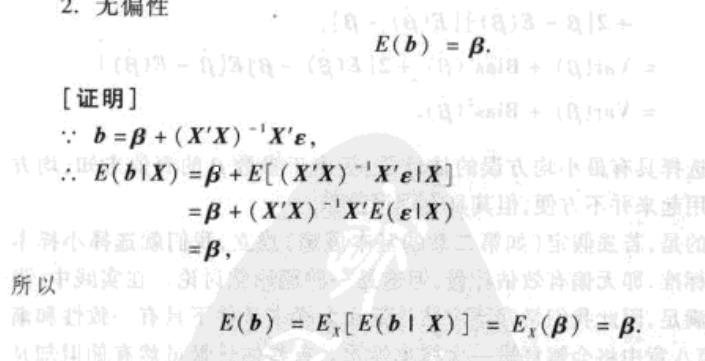
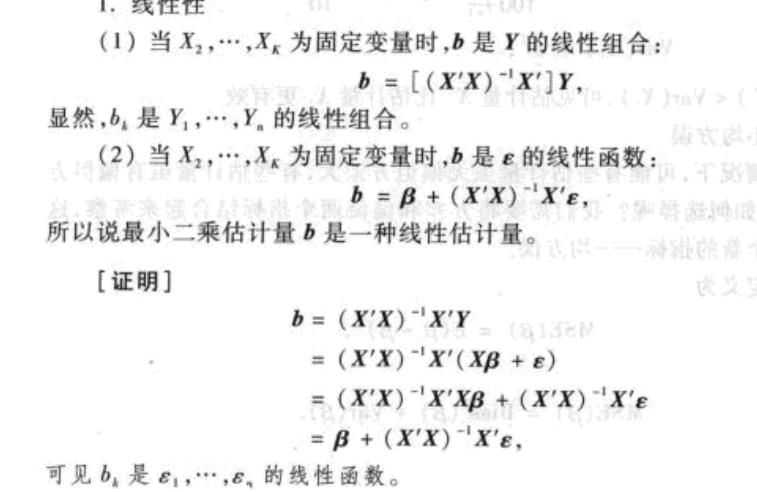
无偏性证明：见伍德里奇中文版（人民大学出版社） 74页和106页

线性性证明： 的一个估计量是线性的充分必要条件见96页

C:\Users\Administrator\Desktop\360截图20160401205948537.jpg即要求能表示成因变量的一个线性函数。

最优性（最小方差性）证明：书本108页

矩阵形式证明参见靳云汇（北京大学出版社）50-51



1. 回归分析中内生性的含义？内生性的来源？解决内生性的方法？

(1) 内生性就是模型中的一个或多个解释变量与随机扰动项相关（Cov[X,u] != 0），这样会使得估计的结果有偏即

\* b = inv(X'X)\*X'y

\* E[b] = E[inv(X'X)\*X'(X\*beta+u)]

\* = E[inv(X'X)\*X'X\*beta] + E[inv(X'X)\*X'u]

\* = beta + E[inv(X'X)\*X'u]

\* != beta

（2）内生性的来源

遗漏变量(omitted-variable bias)→**多重共线性**与解释变量之前的**相关系数**不为0的区别

\* 假设 y = a + x1\*b1 + x2\*b2 + e1

\* Corr(x1,x2) != 0

\* 若估计中遗漏了x2, 则估计模型为：

\* y = a + x1\*b1 + e2 （e2=e1+x2\*b2）

\* 显然，Corr(x1,e2) != 0

测量偏误(measurement error in the regressor)

\* 对于模型 y = x0\*b + e

\* 假设 x = x0 + me (x 是 x0 的样本观测值)

\* 即 x0 = x - me

\* 模型的正确形式应为：

\* y = a + x0\*b + e

\* = a + (x - me)\*b + e

\* = a + x\*b + (e - me\*b)

\* = a + x\*b + e2

\* 显然，

\* Corr(x,e2) != 0

解释变量和被解释变量互为因果、互相作用，如自选择(self selection)

（3）内生性的解决方法

代理变量，尤其适用于遗漏变量产生的偏误；

前定变量，用变量的前一期或前几期数据，可以弱化测量偏误问题；

工具变量

\*- 基本思想：

\* y = X\*b + e

\* 找到一个变量 Z,使得 Corr(Z,e)=0, Corr(X,Z) -> 1

\*

\*-- IV 视角

\* Z'\*y = Z'\*X\*b + Z'\*e

\* => Z'\*y = Z'\*X\*b\_IV

\* => b\_IV = inv(Z'X)\*Z'y

使用面板数据，如固定效应或双重差分。

准实验方法，如断点回归

三

1（1）ols的几个经典假设：即Five GAUSS-MARKOV assumption,①误差项服从标准正态分布即；②同方差性 homoskedasticity；③非自相关性；④解释变量是非随机的；⑤自变量与误差项相互独立即cov(x,u)=0。

（2）家庭income对saving的影响与户主人力资本状况和家庭结构紧密相关。可以推测，户主年龄越大，其人力资本累积程度越高（家庭收入会提高），其家庭结构（如孩子数量）也越趋于稳定。如果模型中删失户主年龄在25岁以下的家庭，该样本选择会造成有偏估计。

（3）假定年储蓄超过25000美元的家庭为富裕家庭，根据消费理论，富人的边际消费倾向低于穷人，由此推测，富人的储蓄选择与穷人也不一致。如果在样本中删失富人群体，那么模型对富人的预测能力将下降，即系数估计出现偏误。

（4）定义：如果抽样不能在各个具有代表性的群体中随机进行，由此会产生样本选择偏误问题。模型的参数估计结果仅适用于该样本，对总体或其他样本缺乏有效的外延解释力。

2（1）Cathhs与误差项有关系：天主教高中多为私立学校，他凭借优越的教育质量，所招收的学生在学习能力等方面优于其他学校，而不可观测的学习能力等与因变量相关，即Cathhs与误差项相关。

（2）题目疑似有误：“大二”疑衍，应为“高二”（前定变量可以更好地作为代理变量）。标准化成绩能充当不可观测的学习能力的代理变量，直接添加进模型，当作控制变量。

（3）cathrel成为有效工具变量的两个条件：相关性 cov（Cathhs，cathrel）≠ 0，要求cathrel（是否为天主教徒）对Cathhs（是否就得天主教高中）中的变异具有相当的解释能力；外生性cov（Cathhs，u）= 0，也称排除性约束，要求cathrel只能通过Cathhs对因变量（是否上大学）发生作用，此外，无其他影响渠道。

可以简单地将cathrel当作因变量对cathrel回归，由决定系数判断工具变量相关性，外生性无法检验。

（4）利用工具变量，可以使用Durbin-Wu-Hausman 检验变量的内生性。

\* H0：模型不存在内生性问题

\*- 基本思想：

\* 若 H0 为真，则 b\_OLS 和 b\_IV 都是一致的，但 b\_OLS 更加有效；

\* 若 H0 为假，则 b\_IV 仍然是一致的，而 b\_OLS 则是有偏的；

简单来说，Hausman比较OLS和IV估计结果是否存在系统性差异，如果存在，则推翻原假设，断言变量为内生。

3（1） 如果非观测效应（unobserved effect）与任何时期的解释变量相关，那么使用固定效应模型或一阶差分。反之，如果独立于所有时期中的每个解释变量，那么当使用随机效应模型。

（2）组内估计量（within estimator）即固定效应估计量，他得之于y和x进行组内变换（within transformation），除去时间均值后，再进行混合OLS估计的结果。组间估计量（between estimator）得之于使用y和x的时间平均值做横截面OLS估计的结果。

（3）固定效应模型利用组内变换，将不随时间改变的非观测效应剔除，这一定程度上控制了遗漏变量导致的内生性问题。

（4）二者（添加20个省市的虚拟变量模型的参数估计结果与固定效应模型估计结果）恰好一样。

4.（1）当p=1时，参与工程劳动力的收入模型为：



可得，工程实施后的平均收入





则工程实施前后的平均变化为：



（2）当p=0时，非参与工程劳动力的收入模型为：



可得，





则工程实施前后的平均变化为：



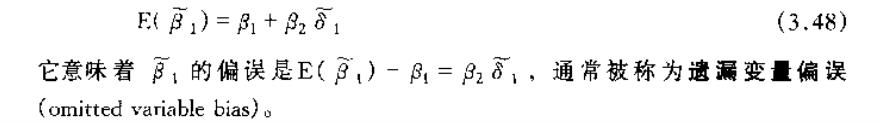
（3）净影响为：

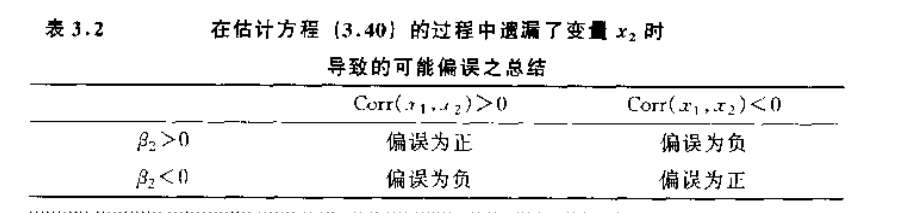
 

（4）参见上文，略。

四、综合题（略）

**押题，遗漏变量偏误**

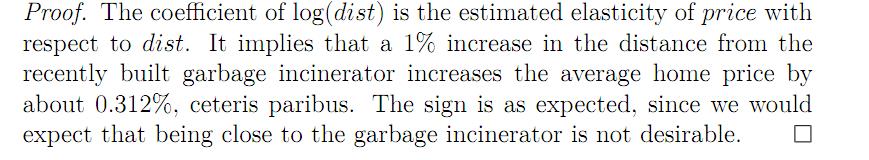




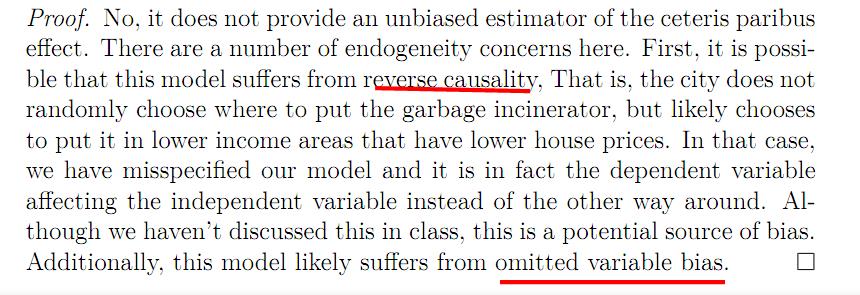
例题The following equation relates housing price (*price*) to the distance from a recently built garbage incinerator (*dist*).



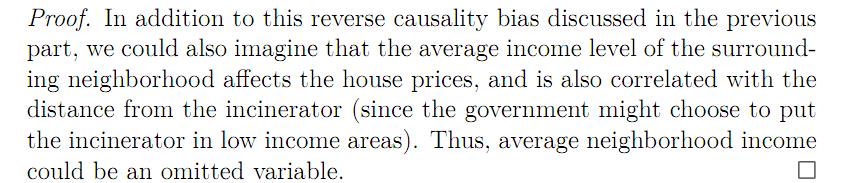
1. Interpret the coefficient on log(*dist*). Is the sign of this estimate what you expect it to be?



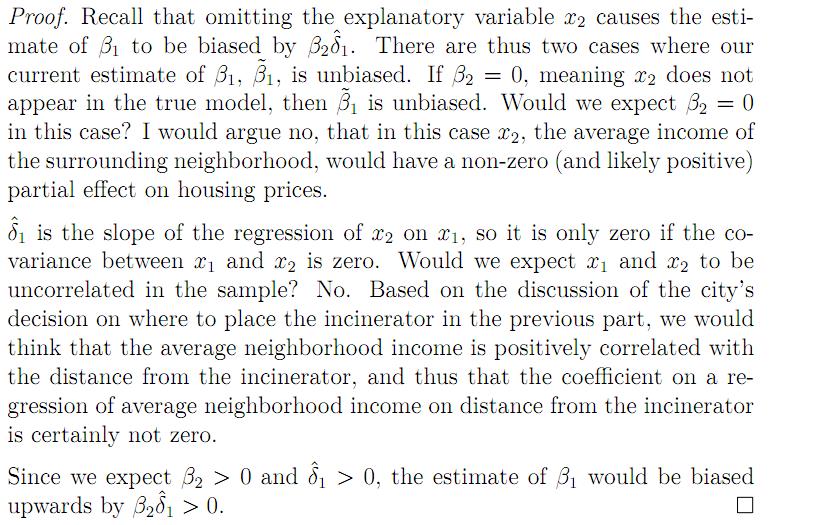
1. Do you think simple regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about how the city decides where to locate its garbage incinerators).



1. What other factors about a house affects its price? Might these be correlated with distance from the incinerator?



1. Choose one of the factors you listed in part c, and explain whether this would bias the coefficient on log(*dist),* and if so, in which direction the coefficient would be biased.



2016/4/2 严如贺 人民大学图书馆 21：00