计量经济学复习资料

Karry

1 计量经济学的性质与经济数据

计量经济学和数理科学最大的区别：主要考虑和分析的是非实验数据。

> 非实验数据(nonexperimental data)：并非从对个人、企业或经济系统中的某些部分的控制实验而得来。(也被成为观测数据(observational data)或回顾数据(retrospective data))
>
> 注意：实验数据仍然存在(Randomized Trial)

1.1 实证检验的过程

Step 1 经验分析 or 规范分析找变量

> 其实在如今实际过程中，我们并不会去关心变量怎么来的，而是直接选取计量模型

Step 2 构建计量模型

> u 中包括了不可观测到的因素 + 可以观测到但不好用数字衡量的因素
>
> 计量说白了就是和干扰项不断斗争的过程

Step 3 估计参数

Step 4 假设检验 + 给结论

1.2 数据类型

- 横截面数据：给定时间点上很多个体 Feature 的数据
- 时间序列数据：个体 Feature 在时间轴上变动的数据
- 混合横截面数据：一年的横截面数据不够，为了扩充样本把多年的混合（两年间个体可以不一一对应）
- 面板数据：很多个体的 Feature 在时间轴上分别变动的数据

1.3 因果效应 和 Ceteris paribus

肥料对作物收成的影响+ 测度教育回报 + 执法队城市犯罪活动的影响

- 很多因素观测不到
- 选择性偏差（Selective Bias）

> 框架：
>
> - 第二章从用一个变量去解释另一个变量的简单线性回归模型开始，了解基础（原理 + 假设）
> - 第三章介绍多元回归分析，开始学习用多个变量去解释一个变量，完成Step3
> - 第四章介绍如何对估计出来的结果，做假设检验，完成Step4
> - 第六章往 OLS 的深层基础进行探究
> - 第七章开始介绍虚拟变量
> - 第八章打破同方差假设，探究如何破除异方差

2 起点：简单回归模型

$$y = \beta_0 + \beta_1 x + u$$

> 两个 bug:
>
> - 不同 x 的边际效应相同
> - u 和 x 之间的影响太重大

引入零条件均值

即：$E(u|x) = 0$ 这意味着不论 x 为多少 u 的均值都是零 u 和 x 在这种情况下无关

> case: 在教育回报率的方程中，意味着 E(abli|16) = E(abli|8) 也就是说受过16年教育人的能力和受过8年教育人的能力完全相同

2.1 OLS 求解参数

两种思路

- E(u) = 0 && E(u|x) = E(ux) = 0
- 最小化残差进行优化

几个性质

- $E(\hat{u}) = 0$
- $E(x\hat{u}) = 0$
- （x均值, y均值）在回归线上
- SST = SSE + SSR

拟合优度 R^2

- $R^2 = SSE/SST = 1 - SSR/SST$

> 经典问题：拟合优度并不是越高越好？ 正确。判断估计结果是否准确的标准是高斯马尔科夫定理，即 BLUE（best linear unbiased estimator-无偏、一致、有效）性质是否成立。而 BLUE 性质是否成立与 R^2 大小没有必然联系。如果高斯马尔科夫假设不成立，估计存在偏差，此时 R^2 再大也没有意义。

几点说明

- 调整因变量单位（/10 or /100）
- 让 y、x 不再线性

> y - x
>
> log y - x
>
> y - log x
>
> log y - log x

- 所谓的线性全部都是围绕着参数 β 所进行的

2.2 OLS 无偏性

> 我们始终渴望 OLS 估计出来的参数是无偏的（无偏可以保证我们的估计是准确的）
>
> 这些是最最基础的内容了

SLR.1(线性于参数)

> 参数 β 之于变量是线性的

SLR.2(随机抽样)

> 有一个服从总体模型方程的随机样本 {(x_i, y_i):i = 1, 2, n} 其样本容量为 n

SLR.3(解释变量的样本有变异)

> x 的样本结果级{x_i, i: 1, ... n} 不是完全相同的值

SLR.4 零条件均值

> E(u|x) = 0

从以上四点假设出发证明**OLS**二元回归估计的无偏性——考前需要再看一下**P47**

注意：这个地方的无偏，只能说明我们的估计与样本中的实际值无偏，但样本选的怎么样谁也不知道

2.3 OLS 估计量的方差

再加上下面的一个假设我们就可以求得 β 的方差

SLR.5 同方差假设

> Var(u|x) = σ^2

P51 会求 β 的方差

2.3.1 问题又来了 σ 怎么求呢？

首先弄清楚误差和残差的区别：误差是整体模型中的 u ，残差是估计模型中的 $\hat{u}$

定义的 $\sigma^2 = E(u^2)$ 所以要是有u 那 σ^2 不就有了 但是谁会告诉你 u 呢？

因此我们只能从样本中得到 $\hat{u}$ 然后估计 σ 得到下面的无偏估计量（P54 有证明）

$$\hat{\sigma}^2 = \frac{SSR}{n-2}$$

至此我们就可以根据一个特定的样本来计算 β 、$se(\beta)$（系数的标准差）、R^2 了此后我们会一直延续这个思路

3 向前推进一步：多元线性回归

> 始终在为 E(u|x) = 0 而不懈奋斗 Ceteris Peribus

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

3.1 OLS 估计参数

和简单线性回归本质是相同的，两种思路 + 一种特殊思路

- 从 E(u) = 0 && E(u|x) = 0 出发
- 从 min(u) 出发 （求导）
- 从真正的 排除其他变量影响上出发：先将一个变量对其他所有变量回归后得到该回归的残差，然后将 y 对残差进行回归。

参数的含义：再其他情况不变的条件下，x_i 每变动一单位 y ...

3.2 偏差（要会写具体过程）

$$\widetilde{\beta_1} = \hat{\beta}_1 + \hat{\beta}_2\hat{\delta}1$$

- 其中波浪线代表遗漏变量后的估计值
- \hat_1 代表真实值
- \hat_2 代表遗漏值对 y 的影响
- \delta 代表相关性（正 or 负）

所谓偏大偏小都是 波浪线 相对于 \hat（真实值）来说的，有了上面的讨论我们自然而然地就有了下面的思考

3.2.1 回归模型中包含了无关变量（过度设定了）

对估计值的无偏性没有任何影响（因为是满足四条假设的）但是会影响到OLS的有效性，P90

3.2.2 偏误的具体描述

和上述提到的偏误完全相同，不再赘述

但是谈及遗漏变量偏误更一般的情形：课本 P88 讲述了一个令人惊叹的真相，对于$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ 这个模型，如果遗漏了 x_3 ，并且 x_3 只和 x_2 相关，那它会不会影响到 β_2 呢，貌似不会？实则会的！

3.3 再论 R^2

在回归中多增加一个自变量后，R^2 绝对不会减小，而且通常会毫无道理的增加

3.4 OLS 的无偏性

SLR.1(线性于参数)

> 所有的相关参数 β_i 之于变量是线性的

SLR.2(随机抽样)

> 有一个服从总体模型方程的随机样本 {(x_i, y_i):i = 1, 2, n} 其样本容量为 n

SLR.3(不存在完全共线性)

> 样本集中没有一个自变量是常数，自变量之间也不存在严格的线性关系
>
> - 只是说不存在线性关系，并不是说没有关系
> - 如果达不到这个条件我们就说存在了完全共线性

如何解释完全共线性的 bug？ 很简单：用其他条件不变来解释！

SLR.4 零条件均值

> E(u|x_i) = 0

从以上四点假设出发证明OLS多元回归估计的无偏性——考前需要再看一下P83

注意：这个地方的无偏，只能说明我们的估计与样本中的实际值无偏，但样本选的怎么样谁也不知道

3.5 OLS 估计量的方差

P90 会求 β 的方差 更要理解他的每一项含义

如果说我要研究我的日常多项支出对成绩的影响，如果多项支出之间都有很强的相关性，那我怎么判断每一个指出的影响呢？

总之：始终以我们研究的因果关系为导向，其他的都可以忽略，因为他们有没有意义对我们来说并没有太多的影响。

那我们现在就有个很大的问题了，考虑无偏的话我们就要尽可能多得添加自变量，但考虑方差我们要尽可能仔细地添加自变量，那该怎么做取舍呢？P94

3.5.1 问题又来了 σ 怎么求呢？（P95）

$$\hat{\sigma}^2 = \frac{SSR}{n-k-1}$$

3.6 高斯-马尔可夫定理

> 在假定 MLR.1~MLR.5下（高斯-马尔可夫假设），OLS 方法所估计出来的参数是 BLUE（最优线性无偏估计量）—— 无偏 + 有效

也就是说 G-M 定理表明了 如果 G-M 假设成立，那么我们只需要采用 OLS 进行参数估计就好，因为别的再好地方法都不如 OLS（就算无偏，也不会有效）

4 估计出来这么多东西，真的有意义吗？假设检验

> 基于前面几章的学习，现在任意给我一个样本，我都可以通过 OLS 来估计 BLUE 的参数值。但是他真的站的住脚吗？也就是说如果上帝给了我们答案，让我们检验一下我们估计的值是否准确，我们怎么检验呢。更进一步，我们没办法接触上帝，我们又怎么能说明估计出来的东西在统计学、经济学上是有意义的呢？

4.1 正态性假定

u 独立于解释变量，服从均值为 0 方差为 σ^2 (MLR 6)

六个假定至此完毕——称为经典线性模型假定（CLM），认为 CLM 包含 GM 假定

MLR.6 是最有力量的假设了，因此也最不容易成立，取对数变分布就是指的更接近正态分布 在这个假设的基础上就可以推的 β 的分布了，进而来检验假设

P113 根据正态分布构建 t 分布（其中证明着实有些困难，但是我们能get到其中的原理即可）

一定要永远记得：我们做的 t 检验都是针对整体的参数是否为 0 绝不能傻傻地去检验估计出来地 β 是否为 0

4.2 单侧假设检验 + 双侧假设检验（既可以检验是否为0 也可以检验其他值）

> 会查表，当自由度 > 120 的时候就可以看成正态分布了

最终描述：在xx显著性水平下，某估计量是统计学显著的；或者说某估计量显著异于零

引了一个求 p 值——就是反向求概率呗，找到实实在在的临界值

4.2.1 经济显著与统计显著

- 经济显著：系数大小和符号
- 统计显著：假设检验

4.2.2 很有趣的一个例子在 P131 如何证明两个估计出来的参数相等？

> 构建新变量

4.3 记住 F 统计量的公式 P137 && P141

P145 最后教会了我们如何报告回归结果

5 深入探究，多元回归的细节不可忽视！

5.1 数据的测度单位对 OLS 统计量的影响

- 被解释变量单位的改变：只会让所有的系数、se 均作等倍的缩小。也就是说经济意义，统计显著性不会有任何改变。同样 R^2 也不会有任何改变。
- 解释变量单位的改变：只会让自己的相关指标改变。其余均不变
- 如果变量以对数的形式出现在模型中，只对截距项有影响，对其他无任何影响。

5.2 β 系数

> 有时候我们要看各个解释变量对被解释变量的影响孰轻孰重 —— 标准化一下

只不过是对所有的变量（包括解释变量和被解释变量）做了减均值，除方差的操作，导致估计参数产生相应的变动。

5.3 更多的函数形式

5.3.1 对数函数形式

> 何时取对数？怎么去估计不再赘述

5.3.2 二次函数形式

> 很简单的道理——有极值，有转折点

5.3.3 含有交互作用项的模型

> 一个解释变量的偏效应，受另一个解释变量的影响

5.4 再论 R^2

5.4.1 调整的R^2

> 之前的 R^2 计算公式明显存在偏误，那我们为何不调整一下呢？

> 另外之前说只要狂加变量我们就会让 R^2 增加，这太差劲了

$$\bar{R} = 1 - (1 - R^2)(n-1)/(n-k-1)$$

5.4.2 利用调整的R方在两个非嵌套模型之间进行选择

5.5 再论偏误与过度

> 如果我控制的太多怎么办？控制的太少又该怎么办？

5.6 有了模型怎么做预测呢 —— 考虑好残差

5.6.1 预测置信区间

> 很明显地：估计出来一个模型，给定一组观测值，就能够拿到被解释变量的估计值，那该如何得到一个置信区间呢？也就是说方差该怎么得到呢？

P195 给你不一样的精彩 思想太简单了，只不过我们想不到

> 误差从哪来的呢？P197告诉你

5.6.2 残差有何用？ —— 残差分析

> 残差表明了实际值和估计之间的差距大小，这一点是判断实际值偏大偏小的重要依据。

5.6.3 very important! 如何将因变量(logy) 转化会实际估计量？—— 做好调整

> 见 P201 很重要！！！
>
> 同时我们还要关心一个问题： 含有 logy 模型怎么样才是有效的呢？ R 方如何去求呢？

6 虚拟变量

7 异方差