（说明： A3双面印刷）

班级： 学号： 姓名：

装 订 线

**杭州师范大学经济与管理学院2016-2017学年第2学期期末考试**

**《计量经济学》试卷（A）**

1. 判断（共10分，每题1分）

1．在用时间序列数据做OLS时，结果的无偏性（Unbiased）和一致性(Consistent)都只需要严格外生性（Strictly Exogenous）作为前提。

严格外生对一致性来说是不够的。 （ W）

2. 异方差（Heteroskedasticity）检验结果显著，说明我们应当拒绝存在异方差的初始假设，认为回归是同方差（Homoskedasticity）的。 异方差检验初始假设同方差（W）

3. 单位根（Unit Roots）检验的结果显著，说明我们应当拒绝存在单位根的初始假设。（R）

4. 高斯马尔科夫定理指出，在满足一系列前提假设的情况下，OLS估计量是最优线性无偏估计量（Best Linear Unbiased Estimator），这是说，OLS估计量在所有估计量当中，都是最优的。 最优的“线性无偏估计量” （W ）

5. 在回归时，解释变量越多，解释变量与误差项相关的可能性就越大，回归结果有偏的可能性因而也就越大。解释变量多了，遗漏变量相应减少，有偏可能就可能会减少 （W）

6. 在多元回归时，部分解释变量间的多重共线性（Multicollinearity），虽然不会导致有偏和不一致，但却会增加所有变量系数的标准误。部分变量间的多重共线性，不影响其他变量（W ）

7. 如果存在处理效应异质性（Treatment-effect Heterogeneity），那即使是两个满足所有前提假设的工具变量（Instrumental Variables），给出的结果也可能存在显著差异。 工具变量估计结果是局部的平均处理效应，不同局部可能有区别 （R ）

8. 线性概率模型（Linear Probability Model, LPM）必然存在异方差。 PPT里证明过（R）

9. 如果分组不是随机的，那双重差分模型（Difference-in-difference, DID）模型的估计结果可能是有偏（Biased）而不一致（Inconsistent）的。 可以理解为零条件均值假设被违背 （R ）

10. 在二阶段最小二乘法中，我们可以直接检验二阶段估计的残差与工具变量的相关性，来检验工具变量的外生性。 不可以，如果工具变量内生，那残差就不是误差项的一致估计量 （W ）

1. 选择（共40分，每题2分）
2. 我们可以通过用Hausman检验对比固定效应(Fixed Effect)和随机效应(Random Effect)的回归结果，来判断面板数据回归中随时间不变的不可观测因素与解释变量是否相关，以下论断中，错误的是：C
3. 这个检验的初始假设是两个回归结果不存在显著差异
4. 如果检验的结果是不显著的，那我们~~只能~~最好使用随机效应模型
5. 如果检验的结果是不显著的，那我们~~只能~~最好使用固定效应模型
6. 如果检验的结果是显著的，那我们只能使用固定效应模型  
   Hausman检验只能说明两个模型的估计结果是否存在显著差异。如果FE和RE存在显著差异，那RE很可能是不一致的。反之，两者可能都是一致的。如果两者都一致，那RE更有效。因而FE的差分会减少变量的变化、增加估计误差。
7. 在时间序列数据研究中，如果我们假设解释变量会对被解释变量产生持续很久的影响，就形成了一个无限期分布滞后模型(Infinite Distributed Lag Models)

这个模型有无限多个参数，因而无法直接识别，不过，如果我们对它的参数施以一些限制，它就会变得可识别，比如，我们可以要求。此时，以下关于这个模型的论断中，正确的是： B

1. 此时，
2. 化简后的模型参数个数减少，因而可识别
3. 我们可以直接估计来获得所有参数的一致估计量
4. 以上都是对的  
   化简后的式子如C中所示，它显然违反零条件均值假设

根据以下Stata命令及其执行结果，完成3-5题

use http://fmwww.bc.edu/ec-p/data/wooldridge/barium,clear

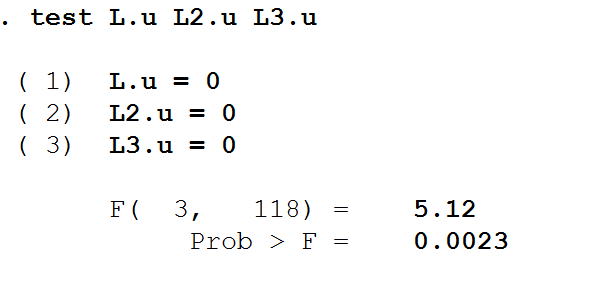
**tsset t**

reg lchnimp lchempi lgas lrtwex befile6 affile6 afdec6

predict u,r

reg u lchempi lgas lrtwex befile6 affile6 afdec6 L(1/3).u

test L.u L2.u L3.u



1. 回归使用的数据是： D
2. 截面数据（Cross Section）
3. 混合截面数据 (Pooled Cross Section)
4. 面板数据 (Panel)
5. 时间序列数据 (Time Series) tsset t。这是设置时间序列结构的命令
6. 这组命令是在检验 C
7. 单位根
8. 异方差
9. **序列相关** (Serial Correlation) 以残差为被解释变量，其滞后项为解释变量。
10. 内生性 (Endogenous)
11. 检验结果说明，要检验的问题 A
12. 存在 显著，说明残差与其滞后项显著相关，这是序列相关存在的证据
13. 不存在
14. 不确定
15. 不知道
16. 关于时间序列中的FGLS，以下论断中错误的是： C
17. 它是为了在序列相关存在的情况下，获得比OLS更有效的估计量
18. 此时的FGLS，实质上是个拟差分（Quasi-Differenced）后的OLS估计
19. 只要OLS是一致的，FGLS就必然是一致的 拟差分的假设要更严格，PPT里有讲。
20. 实践中，这个FGLS经常会反复迭代，直到估计结果收敛为止
21. 如果将回归中的所有解释变量分成两组，分别用和表示**，**那回归式就变成了**，**此时**，**这说明

：B

1. 既可以通过多元回归获得，也可以通过多次一元回归获得 这句话是对的，但无法通过以上推导直接获得
2. 如果，那即使在回归中被遗漏，也不会影响的无偏和一致性
3. 如果，那一旦在回归中被遗漏，就不可能是无偏一致的。
4. 以上都是错的
5. 如果我们想研究受教育程度对收入的影响，我们在也在回归中加入了家庭背景作为控制变量，但因为数据缺失，我们未能在回归中控制智商。如果保持其他条件不变，家庭背景、智商和受教育程度都会提高收入，而家庭背景不会影响智商，但智商和家庭背景的改善都会提高受教育程度，那智商变量的遗漏会导致：C
6. 家庭背景变量的系数不受影响
7. 受教育程度变量的系数依然无偏一致
8. 家庭背景变量的系数会被低估 此时，受教育程度不变情况下，家庭背景的改善意味着智商的下降
9. 受教育程度变量的系数变量的系数会被低估
10. 在面板数据当中，要控制随时间不变的因素有多种方法，我们既可以用LSDV，控制所有个体虚拟变量，也可以用固定效应，将各个变量都减去自身均值在做回归。用哪种方法得到的系数估计值更大？C
11. LSDV大
12. 固定效应大
13. 一样大 LSDV和固定效应的结果毫无二致
14. 视情况而定
15. 以下各种情况，哪个不会导致假设被违背？ C
16. 解释变量存在与真实值无关的估计误差
17. 遗漏了与已有解释变量相关，且对被解释变量也有影响的变量
18. 解释变量对被解释变量的影响存在与其他变量无关的异质性

PPT上有，此时估计结果是无偏而一致的平均影响

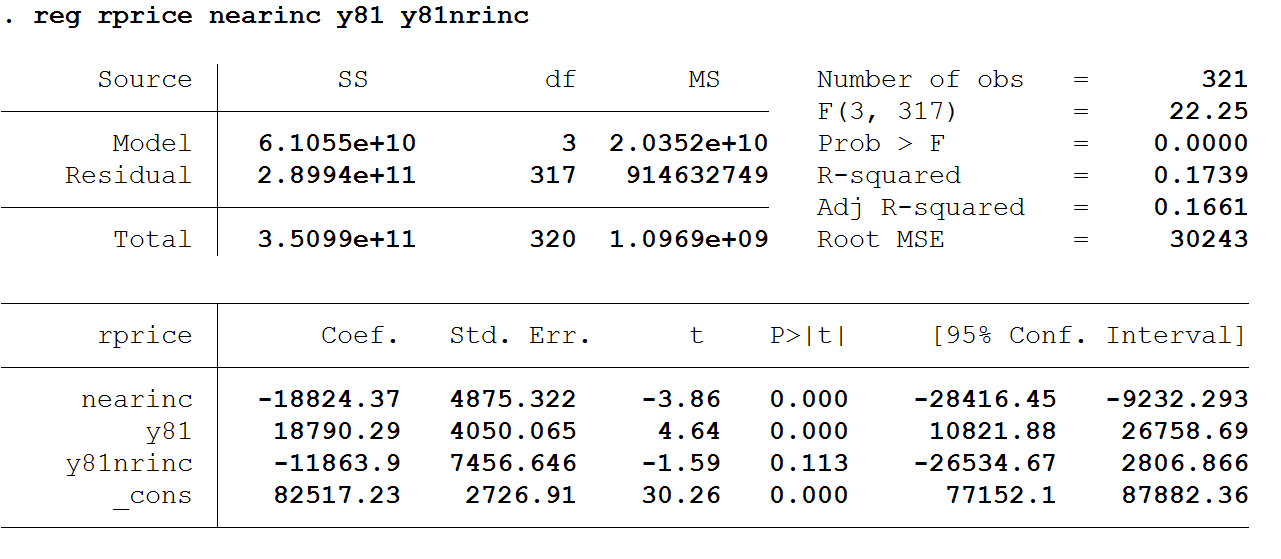
1. 固定效应模型中加入被解释变量的滞后项作为解释变量
2. 以下关于序列相关的论断中，正确的是：D
3. 当序列相关与滞后被解释变量并存时，OLS必然是不一致的
4. 序列相关是时间序列数据的独有问题 面板数据也有
5. 截面数据中不会有序列相关，类似地，时间序列数据不会有异方差
6. 在回归中添加更多被解释变量和解释变量的滞后项，可能可以缓解序列相关 动态完备模型
7. 以下几组命令，哪组的执行结果与其他几组是不一样的 C
8. recode A (nonmissing=1) (missing=0), generate(B)
9. g B=(A!=.)
10. g B=1 if A!=. 当A==.，这个命令会给B赋值为缺失。而其他命令都会给B赋值为0.
11. g B=cond(A==., 0, A/A)
12. 考虑以下联立方程组，,是内生变量，而是外生变量。

，其中哪个方程是可识别的？ B

1. 方程(1)
2. 方程(2)
3. 都可识别
4. 都不可识别
5. 以下关于各种检验的说法，错误的是：A
6. Dickey-Fuller检验可用于检验单位根，其初始假设是不存在单位根

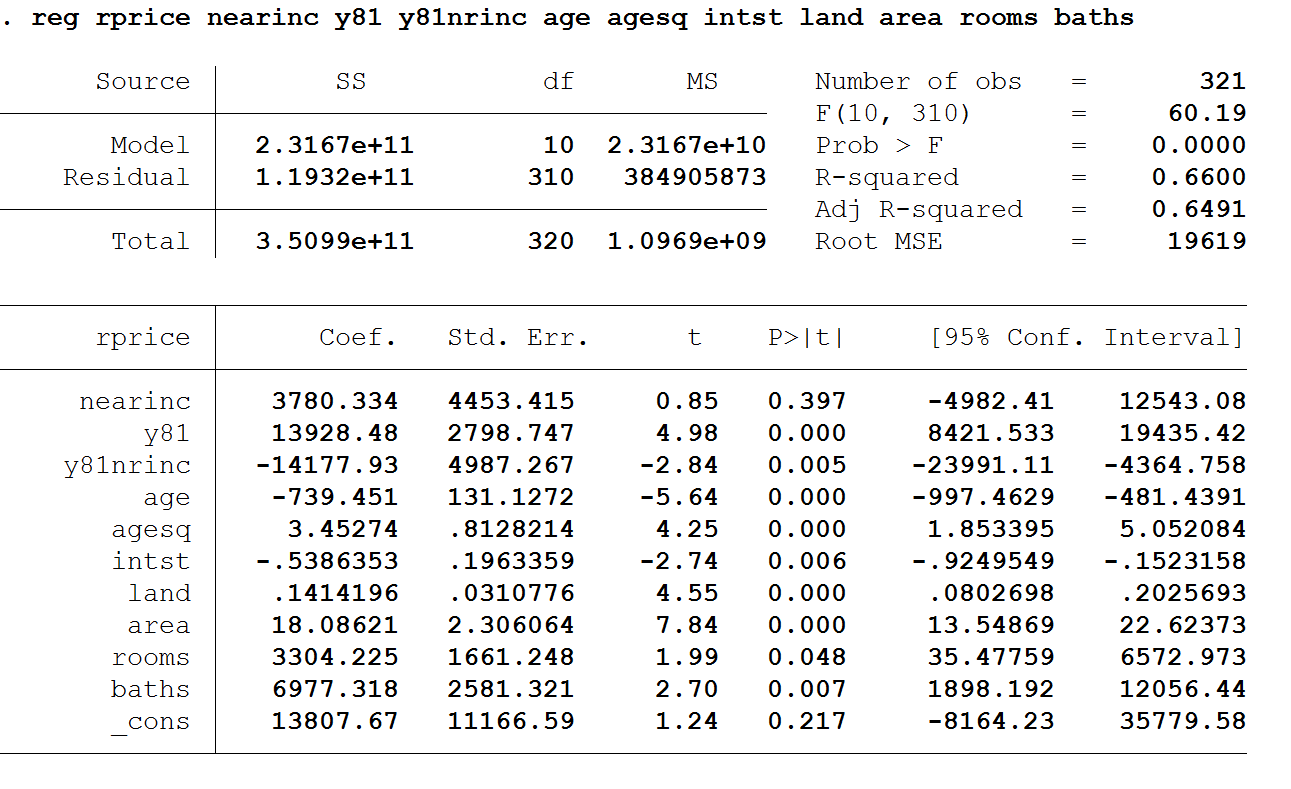
初始假设是存在单位根

1. Hausman检验可用于对比各种假设有差异的回归结果是否存在显著差异
2. 在异方差存在的情况下，一般的t检验和F检验都会失效
3. 小样本情况下，LM检验是无效的
4. 我们可以用双重差分模型来估计处理效应。比如，我们可以把垃圾焚烧炉周边房子视作实验组，而离垃圾焚烧炉较远的房子视作控制组，对比两组房子的价格变化差异，来讨论垃圾焚烧炉建造对房价的影响。如果房价变量是rprice，是否处于垃圾焚烧炉周边变量是nearinc，焚化炉建造于81年，建造前后虚拟变量是y81，两者交叉项是y81nrinc。回归命令和结果如下图所示：



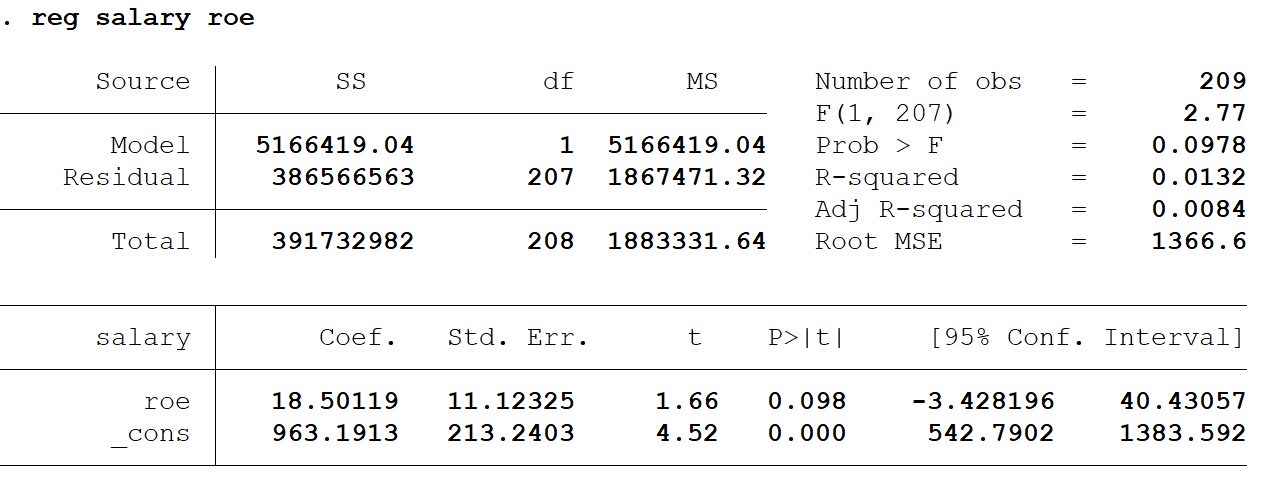
关于这个结果，以下论断中错误的是：D

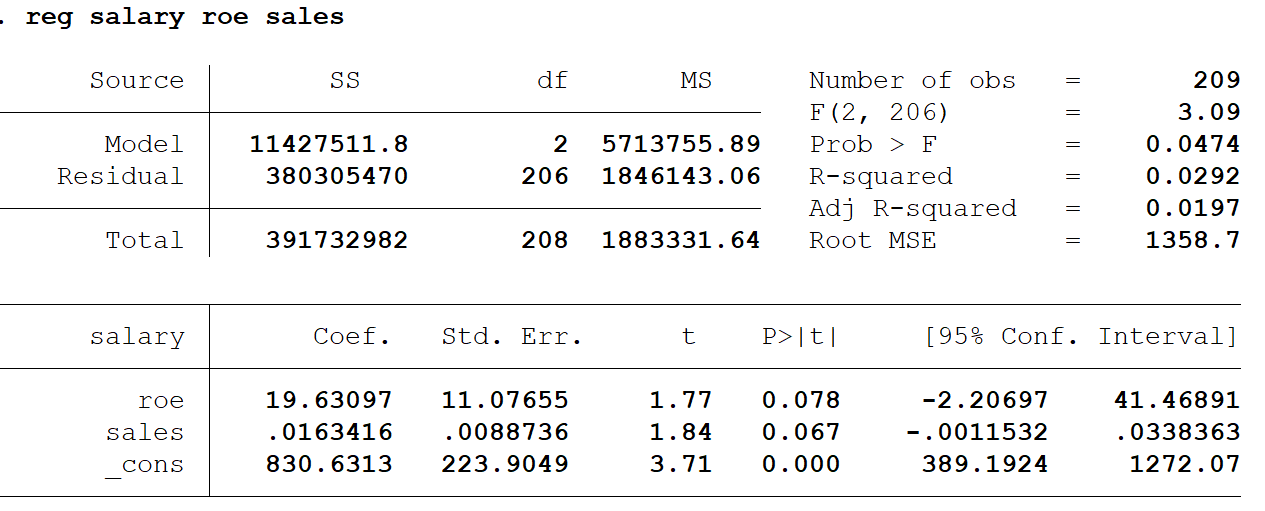
1. 从结果符号看，垃圾焚烧炉建造导致周边地区房价下降了 交叉项系数符号是负的
2. 从显著性看，垃圾焚烧炉建造对周边地区房价的影响是不显著的 交叉项系数不显著
3. 垃圾焚烧炉周边房价相比其他地区要更更低 nearinc系数是负的
4. 以上都是错的 上面三个选项都是对的，所以D是错的。
5. 在15题回归基础上添加更多控制变量后，结果如下图所示：



前后结果的差异说明：B

1. 从显著性看，垃圾焚烧炉建造对周边地区房价的影响是不显著的
2. 距离垃圾焚烧炉远近不同的房子存在显著区别，换言之，焚烧炉选址不是随机的  
   添加这么多控制变量后，交叉项和nearinc的大小和显著性都变量，换言之，这些控制变量与它们是相关的。这是非随机选址的证据。
3. 新添加的控制变量，对解释房价变化并无多少帮助
4. 以上都是错的
5. 如果我们想讨论不同地区在收入的性别歧视上有没有差异，我们样本中有5个地区，那不考虑常数项和其他控制变量，我们需要在回归中加入多少个解释变量？D
6. 4个
7. 5个
8. 8个
9. 9个 2\*5-1
10. 关于匹配（Matching）估计量，以下论断中正确的是 B
11. 它背后的假定与OLS完全一样
12. 基于类似思路，我们可以很容易的获得ATE(Average Treatment Effect)或ATET(Average Treatment Effect on Treated)估计量 PPT上讲过
13. 如果用的是相同数据，针对的是同一问题，它会获得跟OLS完全相同的估计量
14. 以上都是对的
15. 为什么对时间序列，OLS的无偏和一致性并不要求随机抽样，而对于截面数据我们就需要这个假设？ D
16. 因为时间序列数据可能存在自相关
17. 因为时间序列数据可能存在单位根
18. 因为时间序列数据可能存在趋势性和季节性
19. 因为时间序列数据是对同一个体的连续观测，不可能满足随机抽样
20. 以下关于在回归所用的模型中添加一个解释变量的论断中错误的是：D
21. 如果新添加的变量与其他解释变量相关，则会加剧多重共线性问题
22. 是否添加一个变量，其背后应当有经济学理论做支撑
23. 无论新添加什么变量，R2都至少不会减少
24. 我们可以使用逐步回归法，先在回归中加入所有变量，再逐步剔除不显著的，只保留显著的变量。 NEVER逐步回归。
25. 计算及简答题（40分，每题20分）
26. 我们想知道企业CEO的收入是否与业绩相联系，于是我们做了两个回归，都以CEO收入Salary为被解释变量，一个以企业每股收益率（Return On Equity, ROE）为唯一解释变量，另一个则同时加入了企业销售额Sales为另一个解释变量。两个回归的结果如图所示：





1. 完成如下回归结果报告表格（5分）

保留小数点后位数不同不扣分

|  |  |  |
| --- | --- | --- |
|  | 回归1 | 回归2 |
| roe | 18.501  (11.123) | 19.631  (11.077) |
| sales |  | 0.016  (0.009) |
| Adjusted-r2 | 0.0084 | 0.0197 |
| 样本容量 | 209 | 209 |

1. 两个回归中roe变量的系数存在区别，这说明sales变量和roe变量有什么关系？为什么？（3分）负相关

|  |  |  |
| --- | --- | --- |
|  | Corr(x1, x2) > 0 | Corr(x1, x2) < 0 |
| b2 > 0 | Positive bias | Negative bias |
| b2 < 0 | Negative bias | Positive bias |

相比回归2，回归1中roe的系数更小，这是Negative Bias。从回归2看，sales系数>0，也就是b2>0，所以Corr(roe, sales)<0

1. 如果有第三个回归，以sales为唯一解释变量，那其系数相比回归2中系数，会变大还是变小？为什么？（3分）变小  
   因为Corr(roe, sales)<0，而sales和roe的系数都>0，所以忽略掉一个变量会带来Negative Bias，所以系数会变小。  
   这跟回归2中roe的系数更小是一个道理。直接这样类比也能得分。
2. Salary变量和roe变量的相关系数大约是0.1148，如果以roe为被解释变量，以salary为解释变量作回归，其系数大约是多少？（4分）0.0007  
   Corr(X,Y)=,设salary和roe互相回归的两个系数分别是b1,b2，容易证明b1\*b2=Corr(salary,roe)^2。换言之，我们所要求的系数=0.1148\*0.1148/18.501=0.0007
3. 如果用F检验检验sales变量系数=0这个初始假设，得到的F统计量大约是多少？（5分）

回归2是不受约束模型，回归1是受约束模型。F=  
或者，

书上给出了一个更简单的方法，就是回归2中sales变量t统计量的平方。1.84\*1.84=3.39。这个方法上课没讲，写上也对。  
因为小数点保留导致的结果细微差异不扣分

1. 根据所给资料，回答此后问题（20分，每小题5分）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | sex | inc80 | inc81 | inc82 | ue80 | ue81 | ue82 |
| 1 | 0 | 5000 | 5500 | 6000 | 0 | 1 | 0 |
| 2 | 1 | 2000 | 2200 | 3300 | 1 | 0 | 0 |
| 3 | 0 | 3000 | 2000 | 1000 | 0 | 0 | 1 |

上表所示数据，我们可以通过一个简单的命令：

reshape long inc ue, i(id) j(year) 将其转化为如下形式：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| id | year | sex | inc | ue |
| 1 | 80 | 0 | 5000 | 0 |
| 1 | 81 | 0 | 5500 | 1 |
| 1 | 82 | 0 | 6000 | 0 |
| 2 | 80 | 1 | 2000 | 1 |
| 2 | 81 | 1 | 2200 | 0 |
| 2 | 82 | 1 | 3300 | 0 |
| 3 | 80 | 0 | 3000 | 0 |
| 3 | 81 | 0 | 2000 | 0 |
| 3 | 82 | 0 | 1000 | 1 |

1. 如果不用reshape命令，只用generate, replace等简单指令，同样操作要如何完成？（5分）

set obs 9

g year=80 if id!=.

replace id=id[\_n-3] if \_n>3

sort id year

by id: replace year=year[\_n-1]+1 if \_n>1

by id: replace sex=sex[1]

foreach var of varlist sex-ue82{

by id: replace `var'=`var'[1]

}

rename inc80 inc

replace inc=inc81 if year==81

replace inc=inc82 if year==82

drop inc8?

rename ue80 ue

replace ue=ue81 if year==81

replace ue=ue82 if year==82

drop ue8?

以上只是我随手打的一组命令。类似地有无数种实现方式，最笨的方法是一个个变量和数据都自己手动输。

任何能实现最终目的的方法都可以得分。

1. 在转化完成后，输入reg inc ue 命令，得到的ue变量系数是多少？（3分）

-750  
这是个一元回归，一个唯一解释变量是虚拟变量的一元回归。

所以其系数就是ue=1时inc的均值，减去ue=0时inc的均值

也就是(5500+2000+1000)/3-(5000+6000+2200+3300+3000+2000)/6=-750

1. 如果用sex变量给ue变量当IV，输入ivreg inc (ue=sex)命令，最终二阶段结果中，ue变量的系数是多少？（5分）

无意义  
此时，sex是外生变量，ue是内生变量。

IV回归的系数是直接用inc对sex做回归的系数，除以ue对sex做回归的系数。

sex是虚拟变量，因而这两个系数都可以用与上题相同的方法获得。

不过，sex=1和sex=0时ue的均值是一样的，换言之，ue对sex做回归的系数是0，也就是sex对ue没有影响，工具变量与X相关的假设被违背。

因而，回归无法获得有意义的结果。

1. 转化完成的数据是面板数据，在用命令tsset id year设置数据结构后，可以执行固定效应回归的命令，xtreg inc sex ue,fe。获得的结果中，sex变量的系数是多少？（3分）

无意义

这是固定效应回归。而sex是随时间不变的变量，在固定效应回归中会被约掉。

1. 如果运行xi:reg inc sex ue i.id指令，那获得结果与(~~3~~4)中结果有何区别？（4分）

无区别  
这是LLSV，与固定效应无区别。不过，此时sex不一定会被约掉，也可能是某个虚拟变量被约掉。

1. 问答题（10分）

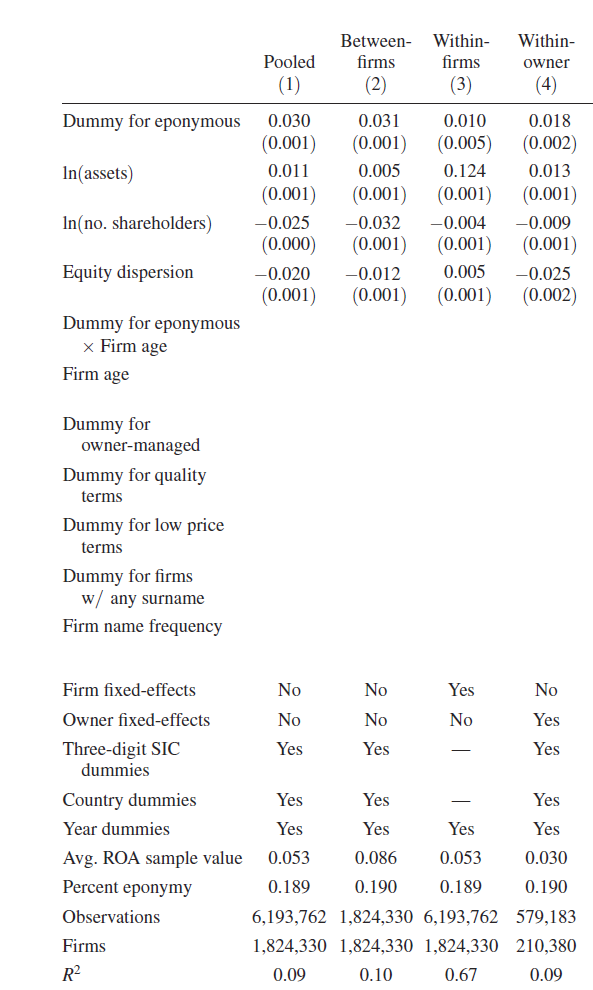
不少企业会以控制人的名字命名。Belenzon, Chatterji和Daley(2017)指出，这一行为可被视作是一个信号博弈（Signaling Game）。控制人以自己名字为企业命名，会提高企业成功对自己的效用，却可能会导致企业的失败给自己带来负效用，因而控制人只会选择用自己的名字给高质量企业命名。

他们用现实数据检验了这一逻辑。具体地，他们以企业资产收益率（Return on Assets, ROA）做被解释变量，以企业名字是否与控制人相同虚拟变量（Dummy for eponymous）为核心解释变量。显然，如果他们的逻辑成立，那我们应该在回归中看到这一虚拟变量的系数显著为正。

他们用的数据来自欧洲，是一个包含超过180万家企业、从2002到2012年近620万个观测值的非平衡面板数据。不少企业背后的控制人是相同的，因而数据中只有不到100万个控制人，其中有接近10%的控制人同时拥有至少两家企业。

此外，有接近20%的企业名字与其控制人名字相同，有5475家公司在样本观测时期内因为更改公司名字或更换控制人，导致企业名字与控制人名字变得相同或者不同。

他们的核心结果包含四个回归。如下表所示，第一列是将所有数据放在一起，做混合OLS的结果；第二列的结果则是先对各个企业的所有变量都在整个样本时期内取均值，将面板数据变成截面数据，再用这一截面数据做OLS；第三、第四列都是面板数据固定效应（Fixed Effects）的结果。通过固定效应，在第三列回归中作者控制了所有随企业变化，而随时间不变的因素。在第四列回归中，作者控制的则是所有随控制人变化，而随时间不变的因素。



这些结果能佐证作者的逻辑吗？对比这~~三~~四个回归的结果，解释它们之间在核心解释变量系数上的差异。

这道题的答案核心不在能不能，而是分析过程。在于为什么这几个回归结果会有差异。

将（Dummy for eponymous）简称为DFE

回归1，作为混为OLS，综合使用了企业间，和同一企业不同时期间DFE的变化来解释ROE的变量。

回归2中所有企业都只有一个观测值，时间上的变化被忽略，因而是在用企业间DFE的变化来解释ROE的变化。

回归1、2DFE系数差不多，意味着DFE对ROE的影响主要出现在企业间。

回归3作为固定效应回归，主要是在用同一企业在不同时期内DFE的变化来解释ROE的变化，换言之，主要在用那些变更了名字或控制人导致DFE变化了的企业在做回归，它的系数要小得多，也不显著的多，显示DFE的影响主要并非出现于这一维度。

回归4也是固定效应回归，但主要是在用同一控制人下不同企业DFE的差异来解释相应的ROE的差异。它的系数介于回归2和回归3直接，显示这个维度上DFE的影响也介于以上两个维度之间。

换言之，不同DFE的企业在ROE上存在变化，主要不是因为换了名字ROE就变好了，也主要不是因为同一老板只把自己的名字给好的企业用，而主要是因为只有好的企业才能使用自己老板的名字。这与作者的逻辑是吻合的。