中国人民大学应用经济学院

博士研究生综合考试样题

（高级计量经济学）

授课教师：徐瑛、谢伦裕、秦萍

参考书目：格林，《计量经济分析》，中国人民大学出版社，2010；Joshua Angris, Jörn-Steffen (Steve) Pischke, Mostly Harmless Econometrics： An Empiricist's Companion, Princeton University Press, 2009

1. 球形干扰假定和 GLS

1) 请写出球形干扰假定，并说明非球形干扰常见情形，及非球形干扰的影响
2) 证明非球形干扰时，GLS 估计量是最优的。

2. 请结合变量选择、测量误差、联立方程、政策评价等情境谈谈你对内生性问题的理解。

3. 作为一个利润最大化的垄断者，你面临需求曲线为：$Q=\alpha+\beta P$。过去，你制定了如下价格并售出相应数量：

Q	3	3	7	6	10	15	16	13	9	15	9	15	12	18	21
P	18	16	17	12	15	15	4	13	11	6	8	10	7	7	7

根据这个数据做 ols 估计，得到： $\hat{Q}=20.7691-0.8408*P$，**且系数估计量的协方差阵为：**

$$\begin{bmatrix} 7.96124 & -0.624559 \\ -0.624559 & 0.0564361 \end{bmatrix}$$

（提示，此协方差阵即为： $\begin{bmatrix} \mathrm{cov}(\hat{\alpha},\hat{\alpha}) & \mathrm{cov}(\hat{\beta},\hat{\alpha}) \\ \mathrm{cov}(\hat{\beta},\hat{\alpha}) & \mathrm{cov}(\hat{\beta},\hat{\beta}) \end{bmatrix}$**），回归标准误为 1.1，$t_{0.025}$=2.16**

假设你的边际成本是 10。根据最小二乘法，

1) 计算利润最大化产出的点估计
2) 计算利润最大化产出期望的 95%置信区间。

4. 极大似然假设 x 服从 Weibull 分布， $f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}, x \geq 0, \alpha, \beta > 0$

1) 求基于 n 个样本随机抽样的对数似然函数

2) 求α，β的极大似然估计方程

5. 我们想研究下一代性别对父母储蓄行为的影响，以验证社会上流行的一个观点：有男孩的家庭因为婚姻市场上竞争所以不得不提高储蓄率。我们将 saverate（储蓄率，被解释变量）对 son（虚拟变量，家庭中是否有男孩），yeduc（户主教育程度），famscale（家庭规模）和 lnwealth（家庭财富对数）进行回归。

1) 讨论在这个回归里为什么 OLS 得出的系数可能是有偏误的。
2) 我们特别纠结 son 这个变量的外生性问题。有文献已经充分证明，第一胎的性别基本上是自然的，而非人为选择的。变量 firstson = 1 指第一胎是否为男孩。讨论 firstson 是否可以作为 son 的工具变量？为什么？

6. 我们想研究注射流感疫苗对大学生成绩的影响，并收集到一些学生的数据：（a）每年是否注射疫苗；（b）每学年的学分绩。疫苗每年 11 月注射，效果发挥到来年 3、4 月。

1) 如果以 $gpa_{it}=\beta_0+\beta_1 shot_{it}+v_{it}$ 做一个 OLS，可能会产生什么问题？举例说明。
2) 这是面板数据，我们可以考虑使用固定效应模型(FE) $gpa_{it}=\beta_0+\beta_1 shot_{it}+a_i+u_{it}$ 。请列出获得无偏性估计的假设。
3) 对于固定效应模型(FE)，写出消除固定效应的表达式。
4) 如果注射疫苗的同学，每年都注射；不注射的同学每年都不注射。请问我们的 FE 研究会遇到什么问题？

7. 本题基于发表在Quarterly Journal of Economics 上的文章"Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program"。该文章探讨的是清理社区的有毒垃圾废弃物(从而改善环境)对于周边房价的影响。在环境质量的变化是来自美国环保局的超级基金计划。超级基金计划的目标是清除美国国家的最严重的危险废物场所。最严重场所的确定包括以下步骤:首先是现场评估，然后Environmental Policy Agency (EPA)的科学家按照Hazardous Ranking System (HRS)给地块打分。如果分值超过28.5，该地块就会被列到"National Priorities List"(NPL)的名录上。随后这些地块上的有毒垃圾会被清理掉。HRS的分值（hrs_82）是在1982年进行的，NPL名录是从1983年开始的。我们关注的变量是地块在2000年为止是否列在NPL名录上（npl2000）。数据所包括的价格（valhs00）是基于在被清理地块2英里范围内的房屋计算出来的。

1) 分别讨论以下几个"事实"如果存在， HRS 的分值可以作为 RD 的有效 running variable 吗？
   a) EPA 说，选择 28.5 作为划分线(threshold)可以使得需要清理的地块数目不太多，在他们清理的能力范围内。
   b) 1982 年参加地块评估的人中，没有任何人知道 28.5 这个切割线。

c) EPA 的文件强调，HRS 的分值并不是很准确。

2) 请问这是一个 sharp RD 还是 fuzzy RD?
3) 描述如何对 povrate(poverty rate)进行分析以验证 RD 的合理性。
4) 描述如何对 hrs 值概率进行分析以验证 RD 的合理性。

8. 《大气污染防治法》规定，根据气象、地形、土壤等自然条件，可以将已经产生、可能产生酸雨的地区或者其他二氧化硫污染严重的地区，划定为酸雨控制区或者二氧化硫污染控制区，即"两控区"。假定从 2000 年起，"两控区"进入实施阶段，"两控区"的控制目标为，到 2010 年：

1) "两控区"内二氧化硫排放量控制在 2000 年排放水平之内；
2) "两控区"内所有城市环境空气二氧化硫浓度都达到国家环境质量标准。

9. 回答以下两题之一

1) 认真阅读这篇文章，Correcting HIV Prevalence Estimates for Survey Nonparticipation Using Heckman-type Selection Models. （Till Barnighausen, Jacob Bor, Speciosa Wandira-Kazibwe, and David Canning，URL: https://www.jstor.org/stable/29764676）。首先弄清楚研究问题，其次写出解决此类问题的实证策略。
2) 请以一个经典随机实验为例，讨论随机实验方法的优势和局限性。

10. 研究人员搜集到 2016 年北京 2 万多户居民的交通出行调查数据，每户居民派一个家庭成员回答问题。调查问卷详细询问了受访居民的个人特征、家庭特征、个人出行方式。

1) 假设个人只有两种方式：公共交通和私家车，如果要估计出行成本和出行时间对出行交通方式的选择，请问什么用的计量模型合适？
2) 如果用线性模型回归，会有什么样的问题？
3) 如果这个被受访者可以有三种出行方式，公共交通、私家车和出租车，针对这样的数据特折，研究人员采用多元 Logit 模型估计，这样的模型估计最主要的缺陷是什么？
4) 第三题模型估计参数和边际影响的区别是什么？

参考答案

==第 1 题==

（1） 球形干扰假定是：$E(uu' / X) = \sigma^2 I$

非求情形干扰情形有两种：异方差性和自相关性。（可适当描述异方差性及自相关性问题）二者虽然对于无偏一致没有影响，但是都产生两个问题（a）标准误计算错误（b）OLS 不再是效率最高（或方差最小）的估计。

（2） 证明：

因为 $\hat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$，其中 $E(uu' / X) = \sigma^2\Omega$

非球形干扰时，$E(uu' / X) = \sigma^2\Omega$，令 $\Omega = C\Lambda C'$，其中 C 的列是 Ω 的特征向量，Λ 为对角阵，对角元素是 Ω 的特征根。

令 $P = C\Lambda^{1/2}$

$\therefore P'P = \Omega^{-1}$ 。

对于 $Y = X\beta + u$，有

$PY = PX\beta + Pu$

记为：$Y_* = X_*\beta + u_*$

且 $E(u_*u_*' / X_*) = P\sigma^2\Omega P' = \sigma^2 I$ 。满足球形干扰假定。所以根据高斯——马尔科夫定理，Y*对 X*做 OLS 回归，所得 OLS 估计量：

$$\hat{\beta} = (X_*'X_*)^{-1}X_*'Y_* = \left(X'P'PX\right)^{-1} X'P'PY = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$$

是最有效率的。

==第 2 题==

内生性问题指 x 和 u 之间存在相关性，小样本下内生性指 E(U/X)=0 不再成立，大样本下，内生性指 cov（U,X）=0 不再成立，内生性问题将导致偏误和不一致。

（1） 变量选择

遗漏变量产生内生性问题。如果遗漏变量既影响 y 又和某个 x 相关，则均值独立假定不再成立，产生内生性问题，系数估计将是有偏不一致的。

（2） 测量误差

令 x1=x1*+e1，其中 x1 是测量值，x1*是真实值，e1 是测量误差。经典变量误差设定下，cov（x1*，e1）=0，那么必有 cov（x1，e1）不等于 0。所以 y 对于 x1 做 OLS 回归，必然违背零条件均值假定，所以必然存在内生性。

（3） 联立方程

如果一个解释变量与因变量被联立决定，那么它通常与误差项相关，这会导致内生性问题。

（4） 政策评价

政策评价中如果对照组和试验组设定非随机，存在自选择问题，则意味着二值变量和 u

存在相关性，产生内生性问题。

==第 3 题==

（1）点估计为：

利润最大化过程意味着 MR=MC，所以利润最大化的 Q 满足 $Q=\alpha/2+5\beta$。所以点估计为

$$\hat{Q}=20.7691/2-0.8408*5=6.181$$

（2）Q 的期望的 95%置信区间求法：

对于利润最大化的 Q，E(Q)= $\alpha/2+5\beta$, $\hat{Q}=\hat{\alpha}+5\hat{\beta}$。

$$P(-t_{0.025}\le\frac{Q-\hat{Q}}{se(Q-\hat{Q})}\le t_{0.025})=95\%$$

$$P(\hat{Q}-t_{0.025}se(\hat{Q})\le Q\le\hat{Q}+t_{0.025}se(\hat{Q}))=95\%$$

而 $se(\hat{Q})=se(\hat{\alpha}+5\hat{\beta})=\sqrt{7.96/4+25*0.056+5(-0.062)}=0.528$

所以区间为[5.04,7.32]

==第 4 题==

1) 求基于 n 个样本随机抽样的对数似然函数

$$\log L = n\log\alpha + n\log\beta + (\beta-1)\sum\log x_i - \alpha\sum x_i^{\beta}$$

2) 求 α，β 的极大似然估计方程

$$\partial\log L/\partial\alpha = n/\alpha - \sum x_i^{\beta}=0$$

$$\partial\log L/\partial\beta = n/\beta + \sum\log x_i - \alpha\sum(\log x_i)x_i^{\beta} = 0$$

（本题中只需要写出方程组，两个方程构成了两个系数的求解方程组，对于无法获得解析解的隐函数，可通过数值模拟求得实际数值）

==第 5 题==

1) 这道题的答题要点是：（1）son 和某个要素 A 相关，（2）A 影响 saverate，（3）A 不在这个模型里，所以 son 和回归的误差相关，导致遗漏变量偏误。
2) 工具变量的两个标准：
   （a）相关性限制（式子）。可以通过 reg son firstson yeduc famscale lnwealth 的回归来验证。
   （b）排除性限制（式子）。讨论如下：如果第一胎的性别基本上是自然的，而非人为选择的，它貌似是外生的变量；但是在中国的计划生育政策下，如果第一胎是女孩，很多农村家庭允许生第二胎，就是不允许生，很多也会生，所以第一胎是男孩的家庭，小孩的总体数量比较少。小孩的数量是会影响储蓄率的，该回归并没有控制小孩数量，所以小孩的数量在误差里，因此 firstson 不满足排除性限制。

==第 6 题==

1) 这道题的答题要点是：（a）shot 和某个要素 A 相关，（b）A 影响 gpa，（c）A 不在这个模型里，所以 shot 和回归的误差相关，导致遗漏变量偏误。
2) 四个假设：
   a) 总体模型线性于参数
   b) 学生个体层面的随机抽样
   c) 解释变量之间不能完全线性相关
   d) $\mathrm{E}(u_{it}|\mathrm{X}, \alpha_i) = 0$
3) 每个个体减去该个体的均值，消去α_i
4) 如果这种情况存在，shot 对于学生个体来说是不随时间变化的，会被固定效应吸收，因此不能估计出 shot 的系数。

==第 7 题==

1) 分别讨论以下几个“事实”如果存在， HRS的分值可以作为 RD 的有效running variable吗？
   a) 以他们清理的能力来设定划分线，显示出划分线并不由于左右两侧地块上所观测到的不同属性（hrs 上的连续变化除外）而设定的，支持 RD 的有效性。相反，如果 EPA 选择 28.5 是因为在 28.5 的附近正好没有地块分布的话，那么分割线两边的地块可能本身就不属于同一个总体，相关属性会很不同，不利于 RD。这一点是可以通过 plot hrs 的概率和 covariate 的均值来间接得到检验的。另外，以清理能力来分割，可以使得列在 NPL 列表上的地块被清理的可能性加大，使得 first-stage 更显著。
   b) 这也会支持 RD 的有效性。即使参加评估的人故意提高某个地块的 hrs 分值从而使相关地块更可能被清理，但因为他不知道这个切割线会在哪里，所以他正好把一个地块从切割线的左边移到右边而且又在我们分析的小区间上的可能性很小。如果他们当中有人知道 28.5 这个界，那些跟“特权”相关联的地块会从 28.5 的左边跳到右边去，这就使得 28.5 的右面积聚了非连续分布的特权地块。谁能保证这样的“特权”不去操作别的政策或者发生别的影响，和清理地块这一行为产生相关性，而又不能被我们控制住。
   c) 这支持 RD 的有效性，只要 hrs 值的偏误不在分割线左右有断裂。如果 hrs 分值不准确到不反应任何有效信息，这就变成一个随机实验的设计。
2) 从图形或者回归可以看出：如果所有28.5以上分数的地块都得到了清理，而所有得分低于该分数的地块都没有得到清理，这是sharp RD。如果有超过该分数的地块没有被清理，而有低于该分数的地块得到了清理，则是fuzzy RD。
3) 将poverty rate作为y变量，地块得分作为x变量，画散点图。并且对数据按照28.5作为分界线分左右两边分开拟合，并检测两边的拟合线在28.5处的截距是否存在统计意义上的显著不同。我们希望看到的是没有显著不同，这样就可以得出结论：28.5左右两边的地块的贫困率并没有显著不同，因此这两边的地块的房价差异并不是poverty rate的不同造成的。
4) 对hrs值概率进行分析，是想看hrs值的分布在28.5左右没有出现显著的断裂。因为如果分布在28.5出现断裂，尤其是略低于28.5的地块显著少，而分数略高于28.5的地块显著多的时候，意味着有很多分数略低于28.5的地块通过某种方法提高自己的分数以进入清理名单。这些跳到清理名单的地块，和还留在略低于28.5分数区域的地块，一定存在某种不可比性，比如跳组地块的政府可能更强势或者更有方法使得该地区获益，那么这些地方的房价高于28.5左边的组，也可能是因为这个原因，而不是这些地方得到了清理的原因。那么RD在这种情况下就是无效的，估计是有偏的。

==第 8 题==

【关键条件】该案例中，假设研究人员明确知道全国那些地级市被划定为“两控区”，也已经计算出所有地级市距“两控区”边界的实际距离。

在【案例背景】下，如果能够找到全国所有城市级 1990-2015 年的二氧化硫历史数据，请设计一个实证策略，尽可能排除其他政策干预。

第一种方法，DID 识别策略。根据案例背景，提取两个关键信息，1）2000 年作为“两控区”政策实施时点，定义政策实施前后的虚拟变量 Aftert（=1，政策实施后，否则=0）；2）被划为“两控区”的地级市作为 Treati（=1，划为两控区，否则=0）。实证回归如下：

$$SO2_{it}=\beta_0+\beta_1\times Treat_i\times After_t+Z_{it}\gamma+\alpha_i+\lambda_t+\varepsilon_{it} \quad (1)$$

其中，Treati×Aftert 可以写成 Treatit，也正确。在地区固定效应 αi 去除不随时间变化的地区固有差异（从而对照组与处理组的组间差异被去除），时间固定效应 λt 去除不随地区变化的时间固有差异（从而政策实施前后差异被去除）之后，研究关注交叉项 Treati×Aftert 估计系数 β1，如果控制变量 Zit 能够排除其他影响政策实施的社会经济因素，那么 β1 可以解释为政策实施对二氧化硫的减排效果。

第二种方法，RDD 识别策略。特别注意：RDD 使用 2000 年以后的数据（政策效果）与使用 2000 年以前的数据（安慰剂检验），结果完全不一样。根据案例背景，提取两个关键信息，1）被划为“两控区”的地级市作为 Treati（=1，划为两控区，否则=0）；2）地级市距“两控区”边界的实际距离 Disti。实证回归如下：

$$SO2_{it}=\beta_0+\beta_1\times Treat_i+f(Dist_i)+Z_{it}\gamma+\lambda_t+\varepsilon_{it} \quad (2)$$

其中，$f(Dist_i)$直接写出 $\beta_2\times Dist_i+\beta_3 Dist_i\times Treat_i$ 或更高次项也正确；同时去掉时间角标 t 和时间固定效应 λ_t 也算正确。RDD 假定距离边界两边越近的地级市，它们的其他社会经济特征就越接近，$SO2_{it}$ 的变化仅仅源于是否被划为“两控区”，而非其他社会经济因素的影响。因此，在 $Dist_i$ 描述了边界距离（从而其他社会经济特征）影响的前提下，研究关注 $Treat_i$ 的估计系数 β_1，解释为被划为“两控区”对二氧化硫的减排效果。

==第 9 题==

（1）答案见文章

（2）开放式论述题）

==第 10 题==

（1）答案：Probit or logit 模型

（2）答案：异方差，以及预测结果概率大于 1 或者小于 0

（3）答案：Independent of Irrelevant Alternatives (IIA)

（4）答案：模型估计参数只能反映 X 变动对 Y 的影响方向，方向和边际影响一致，但估计结果不能反映边际影响。