本试卷适应范围

南京农业大学试题纸

学年　学期　课程类型：必修、选修　试卷类型：A、B

课程号＿＿＿＿＿＿　课程名＿＿＿＿＿＿　学分＿＿＿＿

学号＿＿＿＿＿＿　姓名＿＿＿＿＿＿　班级＿＿＿＿

题号	一	二	三	总分	签名
得分					

一、 简答题（每题 6 分，共 36 分）

1. 什么叫自相关？存在自相关时对 OLS 估计量、t 统计量以及 F 统计量有何影响，如何处理自相关问题？

答：自相关：不同时期误差项存在相关关系(2 分)。

自相关影响：ols 估计量无偏一致，其方差的估计量不再是无偏一致，ols 估计量无效，t 和 F 统计量不服从 t 和 F 分布(2 分)。

自相关处理：用广义可行最小二乘法（PW 估计和 CO 估计）/或者准差分/广义差分法也可以得分(回答出任意一个方法得 2 分)。

2. 工具变量的选择应当具备哪些条件？工具变量回归的基本思路是什么？

答：工具变量的选取有两个基本原则：首先，工具变量必须与内生的随机解释变量高度相关；其次，工具变量必须与误差项不相关，即工具变量本身必须是外生变量 (3分)。

工具变量回归的思路：首先用内生随机解释变量对工具变量以及其他所有外生变量进行回归，得出内生变量的拟合值；然后在原始模型中用拟合值代替内生随机解释变量进行回归(3分)。

3. 什么是模型函数形式误设问题？简要阐述检验函数误设的思路。

答：当一个多元回归模型没有正确地解释因变量和解释变量之间的关系，那就存在函数形式误设问题(3 分)。

检验思路：在回归种加入线性回归预测值得平方项或者更高阶的预测值，并检验相应预测值的系数是否显著为零(3 分)。

4. 什么是测量误差?测量误差产生的原因有哪些?

答:测量误差是指在收集数据过程中登记误差、在数据加工整理过程中的整理误差以及其它统计误差(4 分)。

测量误差产生的原因：

(1) 受人为因素和技术因素的影响，对经济现象和过程的调查登记本身就可能产生误差。

(2) 数据的加工处理过程中也可能导致一定的误差。

(3) 数据的不当使用也会出现误差。（答出任意两点都可得 2 分）

5. 请简述回归元严格外生时 AR（1）序列相关的检验 (每步 2 分)

答：（1）作 y 与 x 的 OLS 回归，得到残差

（2）用残差和滞后一期残差回归，得到系数 $\hat{\rho}$

（3）检验 $\hat{\rho}$ 是否显著

6. 什么是内生样本选择？会带来什么问题？

答：当样本选择以因变量 y 作为基础的样本(3 分)，通常会造成估计有偏(3 分)

二、计算分析题（每题 16 分，共 64 分）

1. 为了估计母亲怀孕期间吸烟对婴儿出生体重的影响，我们构建如下模型：

$$\ln weight = 1.51^{**} + 0.13^{*} cigarrete$$

其中，*weight* 为出生婴儿体重（kg），*cigarrete* 为孕期平均每天吸烟包数，*，**分别表示回归系数在 10%，5%的显著性水平上统计显著不等于 0。

（1）解释回归系数的含义，并分析是否符合预期。

（2）分析此模型是否存在内生性问题，如果存在，列举一个可能造成内生性问题的原因。

（3）如果存在内生性问题，他会给 OLS 估计造成什么影响？

（4）如果我们计划用工具变量法来解决内生性问题造成的偏差，请找到一个工具变量，并讨论其是否满足工具变量的基本假设要求。

参考答案：(每问 4 分)

（1）平均每天多吸一包烟，出生婴儿体重增长 13%。孕期吸烟应该不利于婴儿发育，对婴儿体重有负面影响，本结果不符合预期。

（2）模型存在内生性问题，吸烟数可能跟其他没有控制的影响婴儿体重的因素相关，比如孕妇家庭经济状况，孕妇身体状况等。

（3）内生性会造成 OLS 估计系数有偏且不一致。

（4）可以用香烟的价格作为吸烟数量的工具变量。香烟价格对于单个消费者是确定的，可以认为是外生的，满足第一个条件；同时，香烟价格与吸烟数量应该存在较强的相关性，满足工具变量第二个条件。（答案不唯一）其他工具变量也可以，例如社区吸烟人数的比例等等

2. 在一项关于教育对农村劳动力非农就业影响的研究中，基于江苏、湖北和贵州三省共 1378 个样本得到了如下线性概率模型（LPM）结果：

$$nf_i = 0.64 + 0.17edu_i + 0.20male_i + 0.06age_i - 0.11size_i + 0.18children_i - 0.01finc_i$$

（0.19）（0.05）　（0.10）　　（0.03）　（0.02）　　（0.05）　　（0.01）

其中，*nf* 表示是否从事非农工作（1=是），*edu* 表示受教育年限，*male* 表示性别（1=男性），*age* 表示年龄，*size* 表示家庭耕地面积，*children* 表示家庭儿童数量，*finc* 表示家庭农业收入。

（备注：当观测值数量足够大时，1%，5%，10%对应的双边 t 检验值依次为 2.58，1.96，1.65，括号内为标准误。）

（1）解释受教育年限系数的含义，并检验其显著性。

（2）有研究认为年龄对农村劳动力参与非农就业的影响呈"倒 U 型"，请从理论上解释这种关系；构建一个新的能够描述这种非线性关系的 LPM 模型，并预测年龄变量系数的符号。

（3）在用 LPM 模型结果预测农村劳动力参与非农就业的概率时可能会产生什么问题？有何解决办法？

（4）构建一个模型，使你能检验农村劳动力非农就业是否在江苏、湖北和贵州三省之间存在明显差异。

分）

（2）测量误差对真实的 β_1 和 β_2 的估计影响：$\hat{\beta}_{1(time)} = \hat{\beta}_1, \hat{\beta}_{2(time)} = 3\hat{\beta}_2$ ，所以 β_1 估计无偏，β_2 有偏（5 分）

（3）测量误差对真实的 β_1 和 β_2 的估计影响：β_1 和 β_2 都有偏。（5 分）

4. 如果我们想考察农村老人收入对其健康状况的影响，拟建立以下模型进行实证分析

$$health = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 male + u$$

其中，*health* 表示老人的健康水平，*income* 表示老人的年收入，*age* 和 *male* 分别表示年龄和性别。

（1）如果有人认为子女照料的时间长短（time）对老人的健康也有一定影响，那么上述模型在估计 income 对 health 的影响时会面临什么样的问题？（提示：在农村子女外出务工的背景下，老人的收入相当一部分来自子女外出务工带来的家庭汇款）。

（2）变量 migrant 表示子女外出务工的人数，为什么说它是变量 time 的一个合适的代理变量？

（3）下表包含了有和没有 migrant 作为解释变量时的 OLS 估计值，括号中的数值为标准误。请问为什么第二个模型的 R^2 增加了？

（4）为什么第（2）列中 income 的系数比第（1）系数大？

自变量	(1)	(2)
income	4.37(0.33)	9.49(0.46)
age	-0.75(0.01)	-0.83(0.21)
male	1.28(0.32)	0.77(0.09)
migrant	-	-0.75(0.27)
截距项	9.37(2.11)	8.69(0.68)
R^2	0.04	0.29

参考答案：（每问 4 分）

（1）由于子女外出务工给老人带了一定的经济收入同时也很有可能减少对老人的照料，所以很有可能会造成由遗漏变量造成的估计有偏问题。

（2）因为外出务工人数越多，对老人的照料时间可能减少，两者相关，所以 migrant 可能是一个较好的代理变量。

（3）因为第二个模型中控制了更多的变量，模型解释能力得到了提高。

（4）因为 migrant 和 time 呈负相关，time 本身和 health 呈正相关，在遗漏 time 变量的情况下，income 的估计偏误是负值，因此第（1）个模型 income 的系数会低估。（如果本题仅回答代理变量对模型的结果有纠正作用可以得一半分数。）

参考答案：(每问 4 分)

（1）保持其他因素不变，受教育年限每增加 1 年，参与非农就业的概率增加 0.17。

$t = 0.17/0.05 = 3.4 > 2.58$，表明受教育年限系数在 1%的水平下显著。

（2）经济学含义：当农村劳动力处于青壮年时，更可能有能力和期望从事非农工作来获取更高收入。但农村劳动力处于老年时，随着年龄增长，由于自身健康等因素从事非农工作的可能性越小。

模型：$nf_i = \beta_0 + \beta_1 edu_i + \beta_2 male_i + \beta_3 age_i + \beta_4 age_i^2 + \beta_5 size_i + \beta_6 children_i + \beta_7 finc_i + \mu_i$

即在原始模型中增加年龄的二次型即可。预期年龄的一次项系数为正，二次项系数为负。

（3）用 LPM 模型得到的概率预测值可能出现大于 1 或小于 0 的情况，因此可考虑选择 Logit 模型或 Probit 模型。

（4）在原始模型中加入 2 个省份虚拟变量，例如：

$$nf_i = \beta_0 + \beta_1 edu_i + \beta_2 male_i + \beta_3 age_i + \beta_4 size_i + \beta_5 children_i + \beta_6 finc_i + \delta_1 d_1 + \delta_2 d_2 + \mu_i$$

其中，$d_1 = \begin{cases} 1, & \text{江苏} \\ 0, & \text{其他} \end{cases}$，$d_2 = \begin{cases} 1, & \text{湖北} \\ 0, & \text{其他} \end{cases}$

3. 考虑模型: $y_i = \beta_1 + \beta_2 x_i^* + u_i$

而我们实际上用来度量的 x_i^* 的是这样的 x_i：

(1). $x_i = x_i^* + 5$

(2). $x_i = 3x_i^*$

(3). $x_i = x_i^* + \varepsilon$

其中 ε 是具有通常性质的一个纯随机项

这些测量误差对真实的 β_1 和 β_2 的估计有什么影响.

参考答案：

考虑模型: $y_i = \beta_1 + \beta_2 x_i^* + u_i$

而我们实际上用来度量 x_i^* 的是这样的 x_i：

(1). $x_i = x_i^* + 5$

(2). $x_i = 3x_i^*$

(3). $x_i = x_i^* + \varepsilon$

其中 ε 是具有通常性质的一个纯随机项

（1）测量误差对真实的 β_1 和 β_2 的估计影响：$\hat{\beta}_{1(new)} = \hat{\beta}_1 + 5\hat{\beta}_2$、$\hat{\beta}_{2(new)} = \hat{\beta}_2$，所以 β_1 估计有偏，β_2 无偏（6