计量 HW1

罗淦 2200013522

2024 年 10 月 6 日

1 HW 1

1.1 Classifying Economic Datasets

解答. (a) 数据A是合并截面数据. 因为数据首先在三个不同年份收集, 所以不是截面数据. 因为年份不是连续的, 所以不是时间序列数据. 因为每次收集数据的对象并不是一样的, 因此不是面板数据. 因此是多个截面合并的数据.
(b) 数据B是时间序列数据. 因为在多个年份收集, 所以不是截面数据. 因为是连续的年份上对多个国家的经济数据的记录, 所以是时间序列数据. 但是很难说清楚为什么这个数据不是面板数据, 我的一个解释是: 这个数据更偏向宏观层面, 更适合作为时间序列数据
(c) 数据C是合并截面数据. 因为是每一个截面的抽样是不同的. (d) 数据D是面板数据. 因为是对同一组对象长时间跟踪收集数据. 是对同一批个体在不同时间点上的观察.
数据D和数据A的区别在: D是对同一批个体的记录, 而A的每一个截面的抽样是不一样的.
(d) 数据E是面板数据. 因为跟踪的公司是同一批公司. □

1.2 Log Wage Equation and Returns to Education

解答. (a) $\hat{\beta}_1$的经济学含义: 在固定其他因素不变的情况下, 教育时长增长一年, 每小时的工资在平均意义下增长10%
更喜欢使用log wage equation的原因:
1. 这样得到的回归系数$\hat{\beta}_1$可以直接解释为工资的百分比变化, 比绝对值的变化更方便
2. 工资水平的变化一般会随着工资本身的绝对值的提高而表现出更大的波动, 即会出现异方差问题, 因此取对数可以减少异方差问题
3. 避免解释工资为负数的情况

(b) 如果想要β_1有着教育对工资的因果效应, 我们需要:
1. 零条件均值: $E(u|educ) = 0$. 即解释变量educ不能包含关于u的均值的任何信息, 也即$educ$不是内生的.
否则, 不可能做到$educ$变化的时候, 其他因素都不变化. 例如: 当$educ$增大的时候, 有可能受教育时间更长的人会更聪明(u中的信息), 更聪明的人工资更高
2. 参数线性模型: 即假设真实模型就是$\log(wage) = \beta_0 + \beta_1 educ + u$
否则, 如果教育和工资的关系(系数)是非线性的, 那么回归出来的结果就会错估因果效应
3. 随机采样: 采样的时候不能有选择偏差. 否则回归出来的系数不能解释因果效应, 例如更高教育水平的人一般会更容易进入高薪行业.
感觉很难分析清楚是否穷尽了所有的假设

(c) 未被观察到的变量:
1. 人的能力: 能力更强的人一般会获得更好的教育, 会有更高的工资. 即人的教育水平(一定程度上)包含人的能力的信息, 违背了假设$E(u|x) = 0$.
2. 人成长的家庭的收入: 更高的家庭收入一般可以提供更好的教育, 即人的教育水平(一定程度上)包含了人的成长的家庭收入的的信息, 违背了假设$E(u|x) = 0$.

3. 人际关系: 有好的人际关系的人一般更容易获得高薪的工作, 且社会网络可能通过影响教育机会来间接影响工资水平, 例如: 某些社会圈子可能更倾向于鼓励高等教育. 违背了假设$E(u|x)=0$ □

1.3 Job Training and Workers' Productivity

解答. (a) 不太可能认为企业为员工提供培训的决定和员工的特征是独立的.
可观测的特征:
1. 员工的教育水平, 更高教育水平的员工可能更容易获得培训的机会
2. 工作经验, 有更丰富工作经验的员工可能需要更少的培训.
不可观测的特征:
1. 个人的学习能力: 一些员工的教育水平可能不高, 但是学习能力强, 这是很难被观测到到
2. 团队协作能力和社交能力: 有更强的社交能力的员工一般更值得投资, 但是这很难被观测到

(b) 除了员工的特征之外, 能够影响员工的生产率的因素有:
1. 员工的工作环境
2. 企业的管理模式和激励制度

(c) 即使产出和培训之间是正相关的关系, 也不能推导出培训使得员工的生产能力提高了, 原因:
1. 相关性不等于因果性: 有可能教育水平更高的员工有更高的产出, 但同时, 教育水平更高的员工也更容易获得培训的机会. 因此很难做到固定其他变量不变, 只改变是否获得培训这一因素, 来查看产出的变化. 即不满足零条件均值的假设, 不能做因果推断.
2. 可能没有随机采样: 例如公司只对特定的员工进行培训, 这违反了随机采样的原则, 不能导出因果效应.

(c) (Population)总体: 全体参与考试的工人; (Sample)样本: 参与考试并被随机抽样调查的工人.
注意: 总体数据:全体参与考试的工人的学习时间和成绩; 样本数据: 参与考试并被随机抽样调查的工人的学习时间和成绩

(d) 零条件均值假设: $E(u|hours)=0$.
u中包含的因素有:
1. 参与考试的人的智力和能力
2. 参与考试的人考试当天的身体状态(有可能感冒发烧)
零条件均值假设失效的情况:
1. 遗漏变量误差: 有一个影响考试成绩的因素, 既和复习时间相关(例如: 有更好能力的人有可能学习得更快乐, 会投入更多的时间来复习), 但是没有被模型控制, 就会进入误差项u, 从而导致$hours$中包含了关于u的均值的信息.

(e) 应该是正数, 因为一般复习时间越长, 考试成绩越好.

(f) 如果$E(u)=0$, 那么$E(score)=\beta_0+\beta_1 E(hours)$, 那么$\beta_0$表示, 没有复习的情况下, 工人的平均考试成绩 □

1.4 Theoretical Deduction

解答. (a) 推导$\hat{\beta}_0,\hat{\beta}_1$:
n个样本, 残差$u_i=y_i-\hat{y}_i=y_i-\hat{\beta}_0-\hat{\beta}_1 x_i$

$$\min_{\hat{\beta}_0,\hat{\beta}_1}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2$$

$$\iff \begin{cases} \frac{\partial}{\partial\hat{\beta}_0}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2=0 & \iff \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \\ \frac{\partial}{\partial\hat{\beta}_1}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2=0 & \iff \sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)=0 \end{cases}$$

$$\iff \hat{\beta}_1=\frac{\sum_{i=1}^{n}x_i y_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$$

(b) 在假设SLR.1 - SLR.4下(不需要同方差假设), 推导上述变量的无偏性
根据线性参数假设, 有:

$$y_i - \bar{y} = \beta_0 + \beta_1 x_i + u_i - (\beta_0 + \beta_1 \bar{x}) = \beta_1 (x_i - \bar{x}) + u_i$$

因此有

$$\begin{aligned} E(\hat{\beta}_1) =& E(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}) = E(\frac{\sum_{i=1}^n \beta_1 (x_i - \bar{x})^2 + \sum_{i=1}^n u_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}) \\ =& \beta_1 + E(\frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}) \end{aligned}$$

根据零条件均值假设, 考虑重期望公式:

$$\begin{aligned} & E(\frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}) = E(E(\frac{\sum_{i=1}^n u_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} | x_1, \cdots, x_n)) \\ =& E(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) E(u_i | x_1, \cdots, x_n)) \\ =& E(\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sum_{i=1}^n (x_i - \bar{x}) E(u_i | x_i)) = 0 \end{aligned}$$

因此:

$$E(\hat{\beta}_1) = \beta_1$$

并且有:

$$\begin{aligned} E(\hat{\beta}_0) =& E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\beta_0 + \beta_1 \bar{x} + \frac{1}{n} \sum_{i=1}^n u_i - \hat{\beta}_1 \bar{x}) \\ =& \beta_0 + \frac{1}{n} \sum_{i=1}^n E(u_i) = \beta_0 + \frac{1}{n} \sum_{i=1}^n E(E(u_i | x_i)) = \beta_0 \end{aligned}$$

随机采样假设, 用于保证样本的随机性
非固定解释样本假设, 用于保证分母不为零 □

题目的注记. 不是很确定自己对重期望公式的应用是否正确.