


- 
- 在前面几章关于多元线性回归模型的讨论中，我们讨论的多为定量变量，例如：
 - 工资、受教育的时间、股票收益率…
 - 本章将讨论含有定性信息的变量，例如：
 - 性别、婚姻、种族…

Chapter 7

含有定性信息的多元 回归分析：二值（或 虚拟）变量



章节框架

- 在这一章中，我们将探讨模型中含有定性自变量或因变量的情况
- 首先，我们介绍描述定性信息的方法
- 之后，我们讨论包含定性自变量的情况
- 最后，我们讨论定性因变量的一种特殊情况

对定性信息的描述

- 定性信息

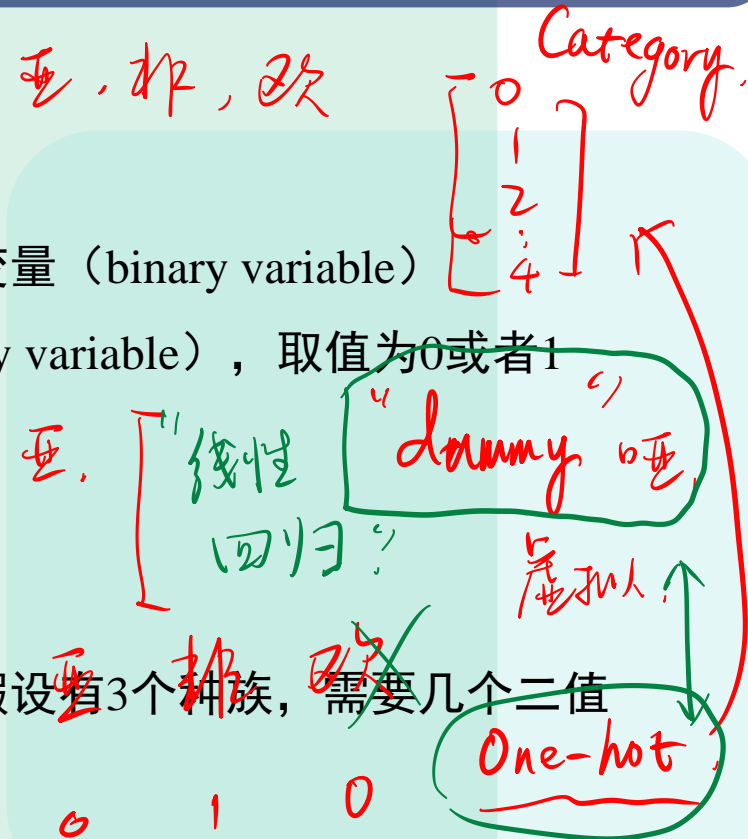
- 例子：性别、种族、行业、区域...
- 整合定性信息的一种方法是使用二值变量（binary variable）
- 最常见的二值变量是虚拟变量（dummy variable），取值为0或者1
- 例子：定义

female = 1, 如果观测对象为女性；

female = 0, 如果观测对象为男性。

- 思考：利用虚拟变量整合种族信息：假设有3个种族，需要几个二值变量？

思考：为什么用0和1描述定量信息？



只有一个虚拟变量

- 考虑只有一个虚拟变量的多元回归模型：

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

male

female = 0

=如果此人是女性而不是男性，则工资收益/损失（其他条件不变）

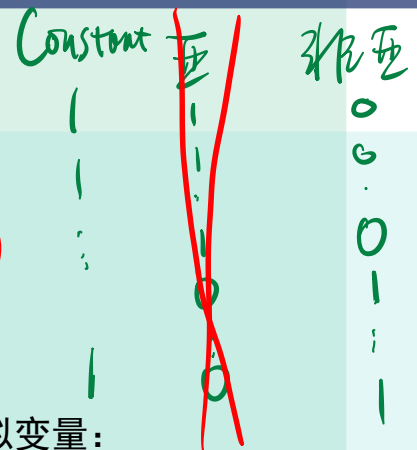
虚拟变量：
=1 如果是女性；
=0 如果是男性

- 零条件均值假定下：

$$E(wage | female = 1, educ) = \beta_0 + \delta_0 + \beta_1 educ$$

$$E(wage | female = 0, educ) = \beta_0 + \beta_1 educ$$

male: 截距.
base: 基准.

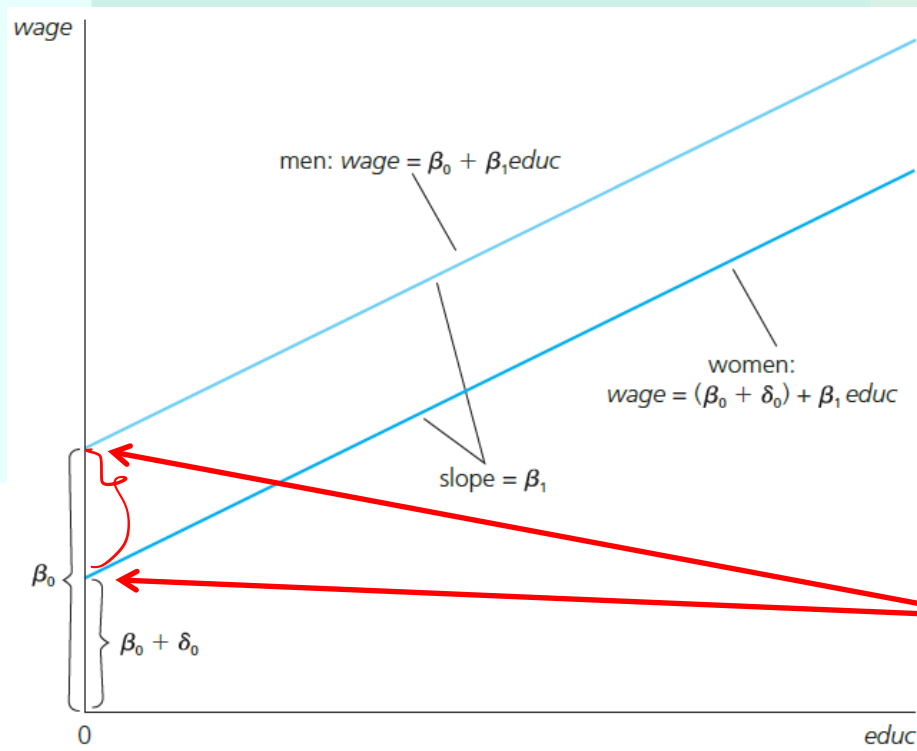


Constant = $\beta_0 + \delta_0$

i.x

只有一个虚拟变量

- 图示



系数的另一种解释：

$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

即，受教育程度相同的男女平均工资的差异

这时，男性样本可以被视为基准组， δ_0 衡量

截距变化

因此，虚拟变量对应系数有更方便的解释

只有一个虚拟变量

- 包含另一个虚拟变量 $male$? 会导致虚拟变量陷阱

此模型无法估计（完全共线）

$$wage = \cancel{\beta_0} + \gamma_0 \boxed{male} + \delta_0 \boxed{female} + \beta_1 educ + u$$

因此，当使用虚拟变量时，总需要省略一个类别：

$$E[u|x] = 0$$

$$wage = \cancel{\beta_0} + \delta_0 \boxed{female} + \beta_1 educ + u \leftarrow \text{基准组是男性}$$

$$wage = \beta_0 + \gamma_0 \boxed{male} + \beta_1 educ + u \leftarrow \text{基准组是女性}$$

或者，可以省略截距项：

$$wage = \gamma_0 \boxed{male} + \delta_0 \boxed{female} + \beta_1 educ + u$$

缺点：

- 1) 更难测试参数之间的差异
- 2) R^2 可能为负值，残差的样本均值不为0（一系列没有截距项回归的问题）

只有一个虚拟变量

- 例子：工资等式

$$\widehat{wage} = -1.57_{(.72)} - 1.81_{(.26)} female + .572_{(.049)} educ \\ + .025_{(.012)} exper + .141_{(.021)} tenure$$

控制，
如果教育、经验和任期固定，女性每小时的收入比男性少1.81美元

$$n = 526, R^2 = .364$$

面板 ← { 时间固定效应，
横截面固定效应。

- 这1.81美元的工资差距不能由男女在受教育程度、工作经历和任期上的差异解释，可能与其他未被控制的生产力特征相关。

只有一个虚拟变量

- 通过虚拟变量比较子群体均值

$$\widehat{wage} = 7.10 - 2.51 \text{ female}$$

(.21) (.26)

$cov(x, u) \neq 0$ (一致小量)



$E(u|x) = 0$ 无偏性

没有控制其他因素的情况下，女性每小时的收入比男性少2.51美元，即男性和女性的平均工资之差为2.51美元。

$$n = 526, R^2 = .116$$

- 讨论
 - 可以很容易地测试均值差异是否显著
 - 与上个模型对比，没有控制教育、经验和任期时，男女之间的工资差距更大；即该模型部分差异是由男女在教育、经验和任期方面的差异造成。

只有一个虚拟变量

- 进一步举例：培训津贴对培训小时数的影响

Policy Evaluation.

每个员工的培训时间

虚拟变量，表明公司是否收到培训津贴

双重差分

$$\widehat{hrsemp} = 46.67 + 26.25 \text{ grant} - 0.98 \log(sales) - 6.07 \log(employ), n = 105, R^2 = .237$$

(43.41) (5.59) (3.54) (3.88)

Program

DID

- 这是一个项目分析的例子

- 处理组 (= 接收补助) vs. 对照组 (= 没有补助)

“双重差分”

只有一个虚拟变量

- 当因变量为 $\log(y)$ 时，对虚拟解释变量系数的解释

$$\widehat{\log(price)} = -1.35 + .168 \log(lotsize) + .707 \log(sqrft) \\ (.65) \quad (.038) \quad (.093)$$

$$+ .027 bdrms + .054 colonial$$

(.029) (.045)

虚拟变量表明房子是否是殖民地建筑风格的

$$n = 88, R^2 = .649$$

$$\Rightarrow \frac{\Delta \log(price)}{\Delta colonial} = \frac{\% \Delta price}{\Delta colonial} = 5.4\%$$

随着虚拟变量从0变成1，房价上涨了5.4%

β 很小时, $e^\beta \approx 1 + \beta$

$\beta \neq \frac{e^\beta - 1}{e^\beta - 1} \neq \beta$

$\frac{y_{new}}{y} = e^\beta$

$(\frac{y_{new}}{y} - 1)$

使用多类别虚拟变量

- 使用多类别虚拟变量

- 1) 通过虚拟变量定义每个类别
- 2) 省略一个类别（使其成为基本类别）

$$\begin{aligned}\widehat{\log(wage)} = & .321 + .213 \text{marrmale} - .198 \text{marrfem} \\ & (.100) \quad (.055) \quad (.058) \\ & - .110 \text{singfem} + .079 \text{educ} + .027 \text{exper} - .00054 \text{exper}^2 \\ & (.056) \quad (.007) \quad (.005) \quad (.00011) \\ & + .029 \text{tenure} - .00053 \text{tenure}^2 \\ & (.007) \quad (.00023)\end{aligned}$$

在其他条件不变的情况下，已婚女性的收入比单身男性低19.8%

$$n = 2,725, R^2 = .0422$$

使用多类别虚拟变量

- 使用虚拟变量来包含序数信息
- 例子：城市信用评级与市政债券利率

市政债券利率

从0-4的信用评分（0=最差，4=最好）

$$MBR = \beta_0 + \beta_1 CR + \text{other factors}$$

Category.

$CR: [0, 1, 2, 3, 4]$

dummy, one-hot.

这种设定不好。更好的方法是定义一组虚拟变量：

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \text{other factors}$$

虚拟变量表明是否选取特定的评级，例如，如果 $CR=1$ ， $CR_1=1$ ，如果不是， $CR_1=0$ 。所有影响都以最差评级作为比较。注意完全共线性。

涉及虚拟变量的交互作用

"DID"

- 允许不同的斜率系数

$$\log(wage) = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot \text{educ} + u$$

Handwritten notes: $\beta_0 + \beta_1 \text{educ}$ (green box), $\beta_0 + \delta_0 + (\beta_1 + \delta_1) \text{educ}$ (green box), $\delta_1 \text{female} \cdot \text{educ}$ (red dashed box, labeled "交互项" with an arrow), $\Rightarrow 0$ under female and $\text{female} \cdot \text{educ}$.

$$\beta_0 = \text{intercept men}$$

$$\beta_1 = \text{slope men}$$

$$\beta_0 + \delta_0 = \text{intercept women}$$

$$\beta_1 + \delta_1 = \text{slope women}$$

- 感兴趣的原假设

$$H_0 : \delta_1 = 0$$

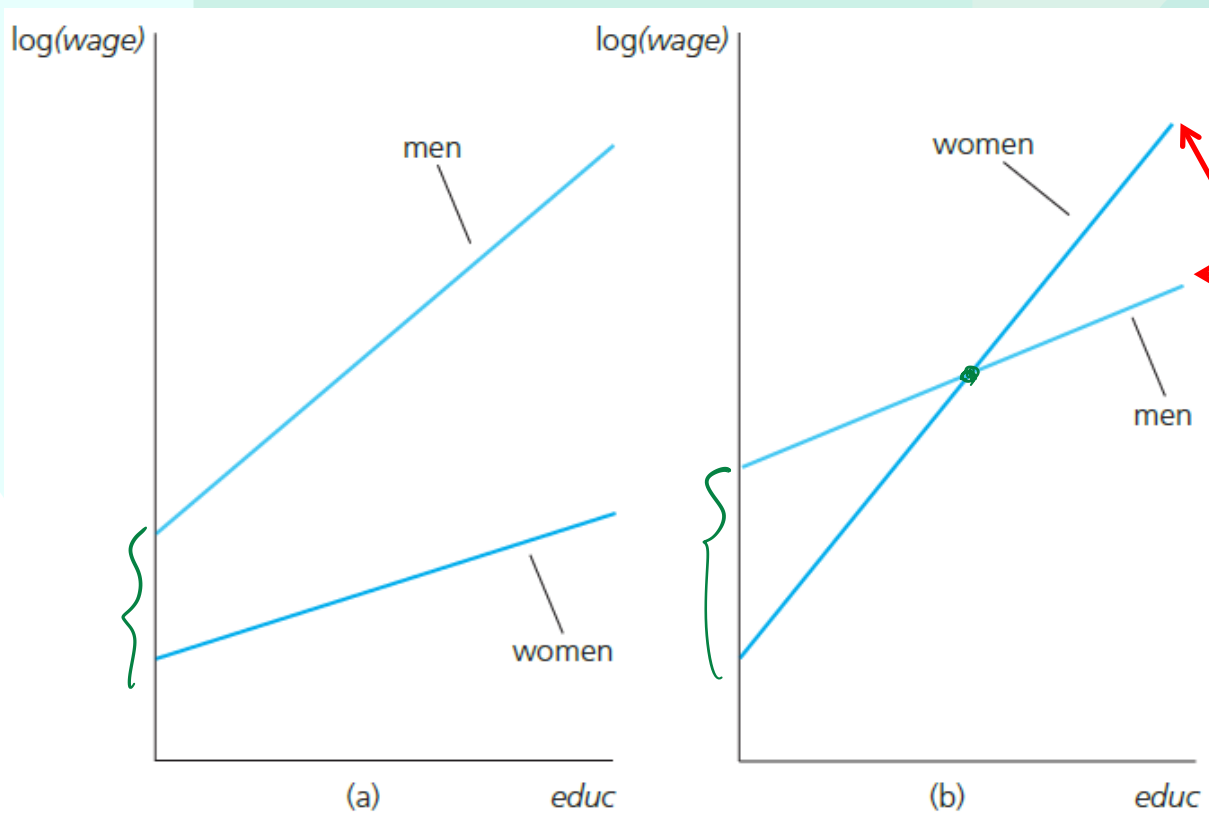
对男性和女性来说，教育的回报是一样的

$$H_0 : \delta_0 = 0, \delta_1 = 0$$

对于男性和女性来说，整个工资等式是一样的

涉及虚拟变量的交互作用

- 图示



表明使用女性虚拟变量的截距和斜率
能够为男性和女性建立完全独立的工
资方程

涉及虚拟变量的交互作用

- 对数小时工资方程

$$\widehat{\log(wage)} = .389 - .227 \text{ female} - .082 \text{ educ} \\ (.119) \quad (.168) \quad (.008) \\ - .0056 \text{ female} \cdot \text{educ} + .029 \text{ exper} - .00058 \text{ exper}^2 \\ (.0131) \quad (.005) \quad (.00011) \\ + .032 \text{ tenure} - .00059 \text{ tenure}^2, n = 526, R^2 = .441 \\ (.007) \quad (.00024)$$

Handwritten notes: $\frac{1}{(1-R_x^2)}$ above the female coefficient; a green arrow points from the female coefficient to the interaction term; a red arrow points from the interaction term to the text below.

没有证据拒绝男性和女性的教育回报率是相同的原假设

这是否意味着没有明显的证据表明，在同等学历、经验和终身职位的女性薪酬较低？

$$\boxed{\text{female} \cdot (\text{educ} - 12.5)}$$

C7.
“断点回归”

涉及虚拟变量的交互作用

- 检验不同组之间回归函数上的差别
- 例子：考虑女生和男生GPA回归函数是否有差别？

机制检验

同方差

大学GPA

SAT分数

高中排名百分比

$$\begin{aligned} cumgpa = & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female sat + \beta_2 hsperc \\ & + \delta_2 female hsperc + \beta_3 tothrs + \delta_3 female tothrs + u \end{aligned}$$

大学课程的总学时数

- 男生组 ($female = 0$) :

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

- 女生组 ($female = 1$) :

$$cumgpa = \beta_0 + \delta_0 + (\beta_1 + \delta_1) sat + (\beta_2 + \delta_2) hsperc + (\beta_3 + \delta_3) tothrs + u$$

全面回归

男生、女生

机制A, B.

涉及虚拟变量的交互作用

- 无约束模型（包含全部交互项）

$$\begin{aligned} cumgpa = & \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc \\ & + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u \end{aligned}$$

- 原假设：

$$H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

所有交互效应为0，即相同的回归系数适用于男性和女性

涉及虚拟变量的交互作用

- 无约束模型的估计

$$\begin{aligned}
 \widehat{cumgpa} = & 1.48 - .353 \text{ female} + .0011 \text{ sat} + .00075 \text{ female} \cdot \text{sat} \\
 & (.21) \quad (.411) \quad (.0002) \quad (.00039) \\
 & - .0085 \text{ hisperc} - .00055 \text{ female} \cdot \text{hisperc} \\
 & (.0014) \quad (.00316) \\
 & + .0023 \text{ tothrs} - .00012 \text{ female} \cdot \text{tothrs} \\
 & (.0009) \quad (.00163)
 \end{aligned}$$

单独测试，每个都不显著

$$n = 366, R^2 = .406, \overline{R^2} = .394$$

$\text{Var}(u|x) = \sigma_u^2$ 同方差.

Chow: A $y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + u$

B. $y = \tilde{\alpha}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{u}$

(A+B). $y = \check{\alpha}_0 + \check{\beta}_1 x_1 + \check{\beta}_2 x_2 + \check{u}$

$(\beta_1 = \tilde{\beta}_1, \beta_2 = \tilde{\beta}_2)$ Restricted model.

涉及虚拟变量的交互作用

- F统计量的联合检验

Restricted

原假设被拒绝

population

$$F = \frac{(SSR_P - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} = \frac{(85.515 - 78.355)/4}{78.355/(366 - 7 - 1)} \approx 8.18$$

- 计算F统计量的替代方法

- 对男性和女性分别进行回归；无约束的SSR由这两个回归的SSR之和给出

- 进行约束模型回归，记录SSR

- 这种方法叫做Chow检验

- 重要条件：原假设下，检测假设各组之间的误差方差相同

1948. Hadvemo.

(1960)

Chow?

Gregory Chow.

Adam Smith

杨小凯 "分工理论"

已知机制转换存在。
A. B 差距。 } 锦上添花

“不知机制转换。” 白聚山. 陈. 晓虹.

Bai & Perron (1998) (2003)



$y = \mu_j + \beta'x + u.$

☆ 检验1: 没有 vs. 有! ✓
 ☆ 检验2: 有1 vs. 有2? ✓
 ✗ 检验3: 有2 vs. 有3! ✗

二值因变量：线性概率模型

- 因变量为二元值的线性回归

$$y: [0, 1]$$

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$$

$E(u|x)$ 如果因变量取值为1和0

$$\Rightarrow E(y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

$$E(y|x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x)$$

线性概率模型 (LPM)

$$\Rightarrow P(y = 1|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

- $P(y = 1|x)$ 又被称为响应概率 (response probability)
- 斜率系数的含义：

$$\Rightarrow \beta_j = \Delta P(y = 1|x) / \Delta x_j$$

在线性概率模型中，系数描述了解释变量对y=1概率的影响。

二值因变量：线性概率模型

- 例子：已婚妇女的劳动力参与

=1 如果是劳动力, =0 如果不是

丈夫收入（单位：千美元/年）

$$\widehat{inlf} = .586 - .0034 \text{ nwifeinc} + .038 \text{ educ} + .039 \text{ exper}$$

(.154) (.0014) (.007) (.006)

$$- .00060 \text{ exper}^2 - .016 \text{ age} - .262 \text{ kidslt6}$$

(.00018) (.002) (.034)

$$+ .0130 \text{ kidsge6}, n = 753, R^2 = .264$$

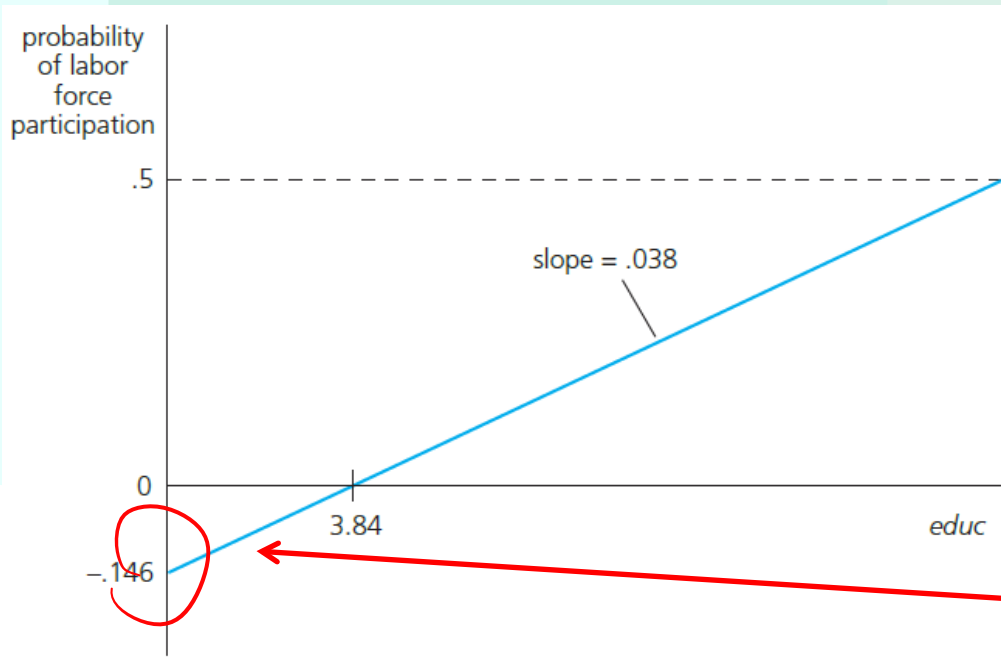
(.0132)

如果六岁以下儿童数量增加一个，
女性工作的概率下降0.262

看起来不显著

二值因变量：线性概率模型

- 例子：已婚妇女的劳动力参与（续）



> 1
 < 0
nwifeinc=50, exper=5, age=30,
kindslt6=1, 并且 kidsge6=0

样本中的最高教育水平为educ=17。
在给定的情况下，这导致劳动力的预测概率约为50%。

预测概率为负，但没有问题，因为样本中没有女性的教育程度低于5年。

二值因变量：线性概率模型

线性概率模型的缺点

- 预测概率可能大于1或小于0
- 边际概率效应有时在逻辑上是不可能的
- 线性概率模型必然是异方差的

$$\text{Var}(y|x) = P(y = 1|x) [1 - P(y = 1|x)]$$

伯努利变量的方差

- 需要计算异方差一致的标准误差

线性概率模型的优势

- 易于估计和解释
- 在实践中，估计的效果和预测通常是可接受的
- 一个更为广泛应用的模型为二值选择（binary choice）模型

$$(1-p)^2 p + (0-p)^2 \cdot (1-p)$$

均值

$y: [0, 1]$

1: 实验组
0: 对照组

Logic, Logistic
Probit 模型

"PSM" 匹配

DID → Propensity Score Matching

对政策分析和项目评价的进一步讨论

- 例子：工作培训津贴对工人生产率的影响

公司的废品率

=1 如果公司接收津贴, =0 不接受

$$\widehat{\log(scrap)} = 4.99 - .052 grant - .455 \log(sales)$$

(4.66) (.431) (.373)

$$+ .639 \log(employ), n = 50, R^2 = .072$$

(.365)

对生产率没有明显影响

处理组：津贴接受者，控制组：不接受津贴者

可能存在的问题：拨款以先到先得的方式发放，而不是随机分发。存在一种可能：工人平均水平（教育水平，能力和工作经历等）较低的公司看到了提高生产力的机会，就先申请。

“抽子 去医院”

→ 实验组、随机 → y, ✓

对政策分析和项目评价的进一步讨论

- 自选择问题

- 在给定的和相关的例子中，处理状态可能与影响结果的其他因素有关
- 如果这些因素被忽略，会造成估计结果的偏误

- 相应的，如果在实验中处理组的分配是随机的。在这种情况下，可以使用简单的回归来推断因果效应

$$y = \beta_0 + \beta_1 \text{partic} + u$$

虚拟变量表明是否进行处理与影响结果的其他因素无关。

对政策分析和项目评价的进一步讨论

- 非白人客户是否受到歧视？

虚拟变量表明是否批准贷款

种族虚拟变量

信用评级

$$\text{approved} = \beta_0 + \beta_1 \text{nonwhite} + \beta_2 \text{income} + \beta_3 \text{wealth} + \beta_4 \text{credrate} + u$$

- 重要的是要控制对贷款批准可能很重要的其他特征（例如职业、失业）
- 忽略与非白人虚拟变量相关的重要特征将会产生估计的偏误

小节

- 我们在本章中了解了如何在回归分析中使用定性信息
- 为了区别两个组，可以定义一个虚拟变量
- 对于 g 个组的情况，需要定义 $g-1$ 个虚拟变量
- 对于虚拟变量估计值的解释，都是相对于基准组而言的
- 虚拟变量与定量变量交互项可以使不同组出现不同的斜率
- 我们可以用F统计量检验组间差异
- 我们讨论了使用二值响应因变量时模型斜率系数的含义