第八讲 离散数据分析

选择困难症的曲线救国

耿瑞霞（嘎嘎）

清华大学公共管理学院

清华大学学生学习与发展指导中心 清华大学公共管理学院党委研究生工作组

清华大学学生学习与发展指导中心
乐学

清华大学公共管理学院党委研究生工作组
公管声音

适合什么人群

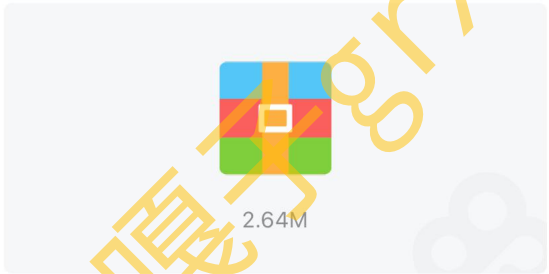
- 0基础：跟我走，妥妥的
- 半瓶子晃荡：
  - 请问Logit能用R^2看拟合优度吗？
  - 请问回归系数有意义吗？怎么解释？
  - 加入IV系数变大还是变小？
  - 使用交互性出现多重共线怎么办？
  - 调节效应、中介效应、交互效应一样吗？
- 高手
  - 在下献丑了多多指教

2小时你能get到什么?

- 1. Why——为啥离了离散模型你的学术生涯就到头了
- 2. What——离散什么鬼?
- 3. Difference
  - probit和logit的世纪之争
  - 几率、几率比、概率
  - 交互效应
- 4. How——

课程资料

club_离散 0311.zip



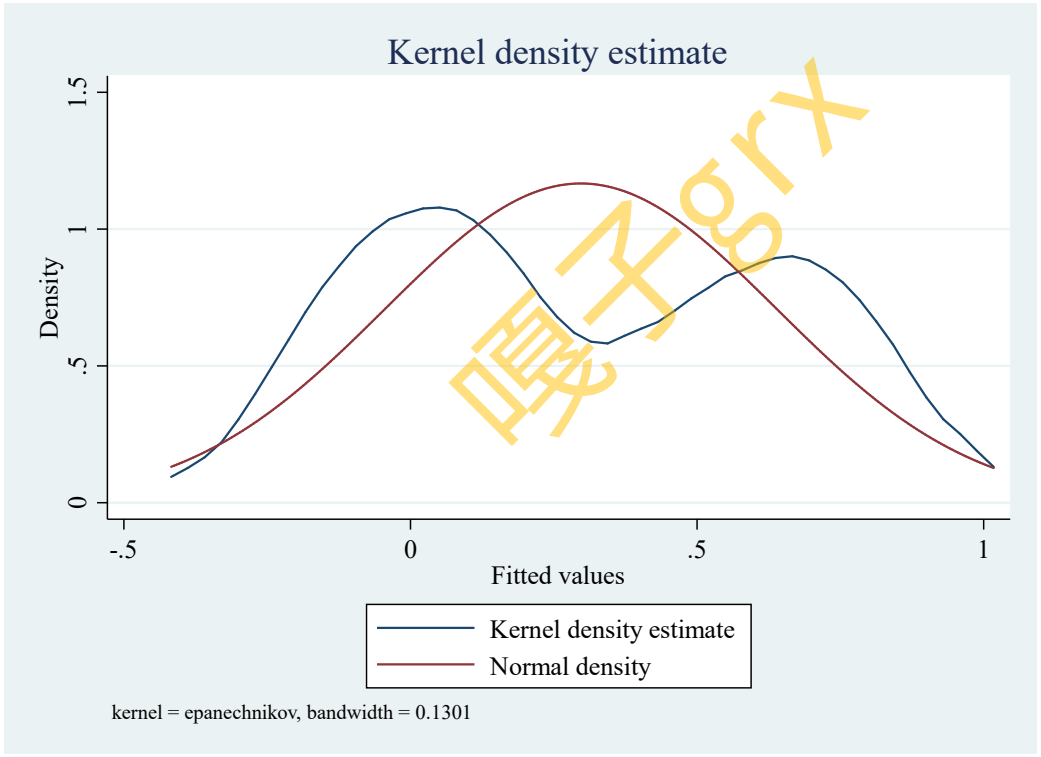
微信长按识别二维码获取文件

有效期至 2020-03-17 10:37:34

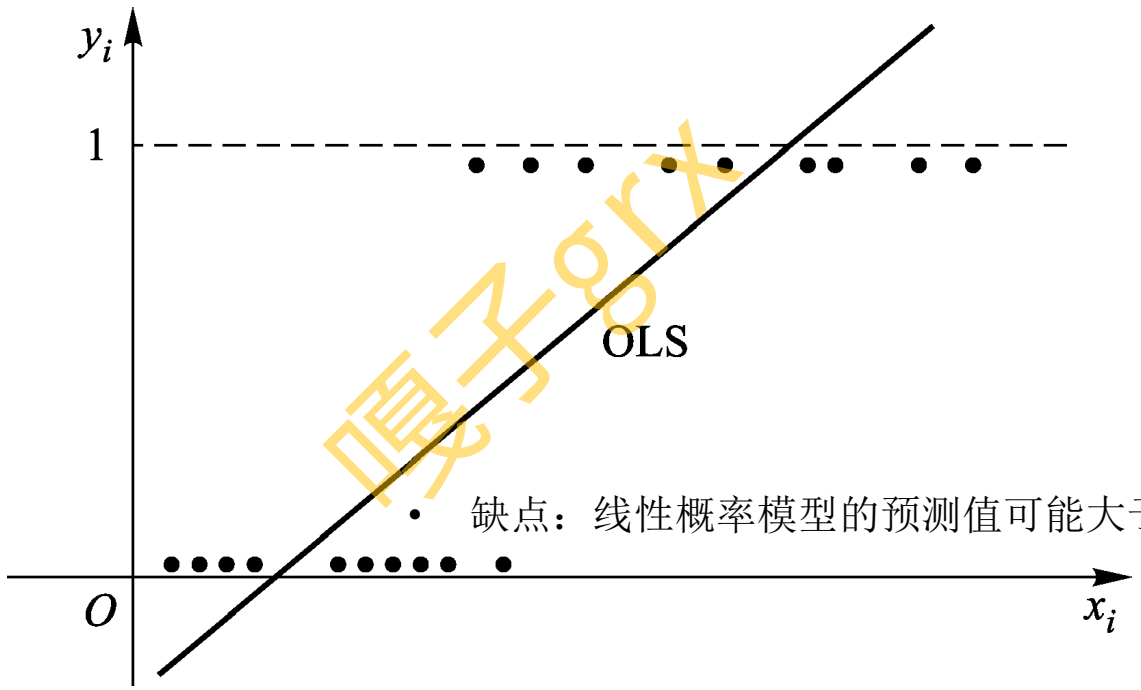
Hamlet's choice: To be or not to be?

- 你跟梦寐以求的女神or男神牵手的几率多大？
- 你们队吃鸡赢面的概率有多大？
- 这个阶段是推塔还是打龙更能一举拿下水晶塔？
- 硕士毕业后是去职场给老板搬砖还是申博换个地方搬砖，还是自己索性做CEO，赢取白富美，走向人生巅峰？

1. Why——为啥离了离散模型你的学术生涯就到头了

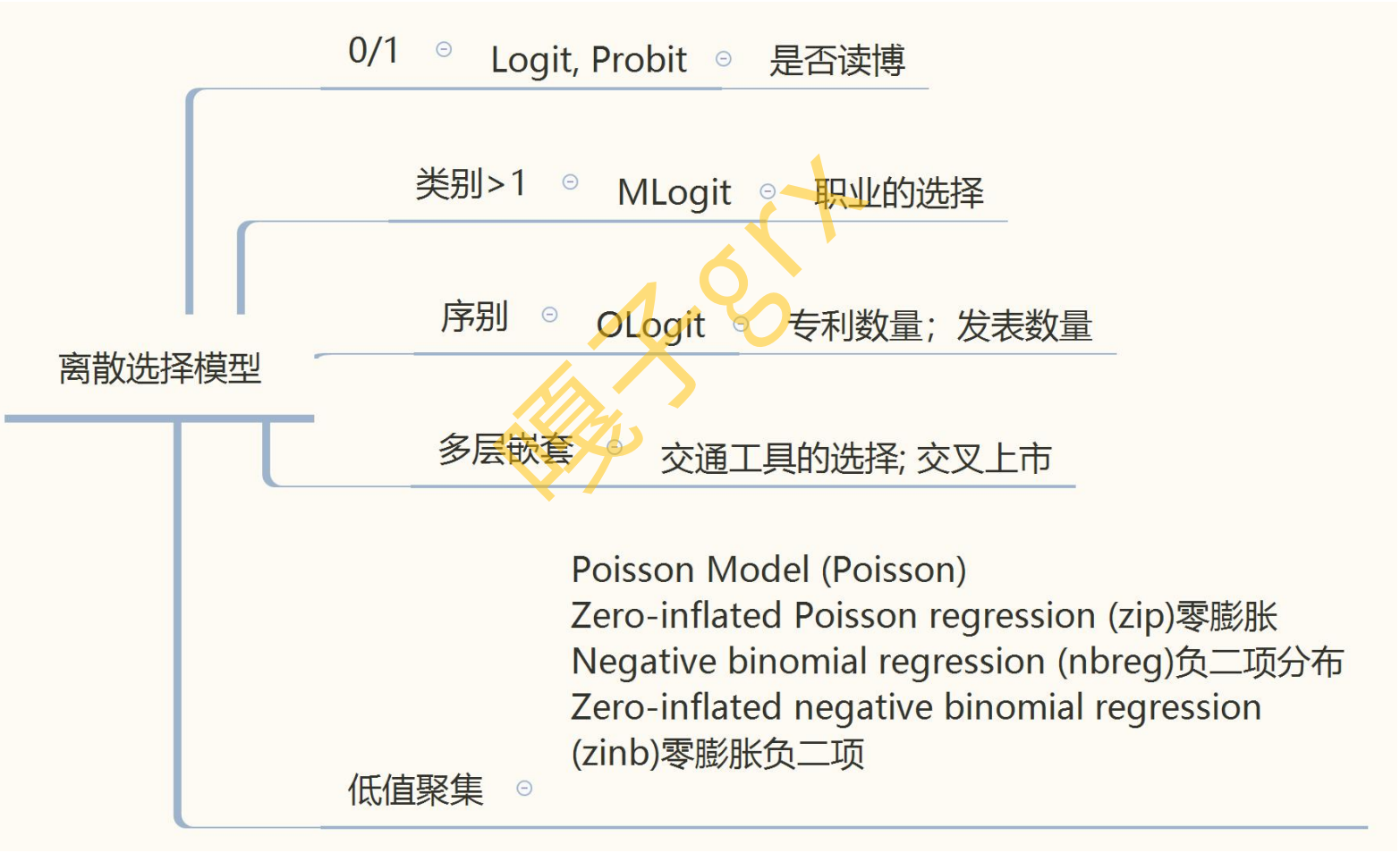


线性概率模型的缺点



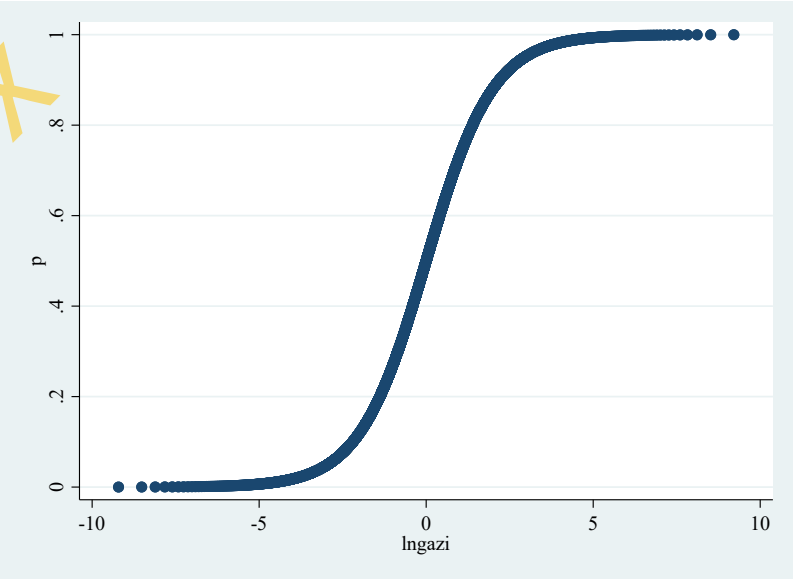
缺点：线性概率模型的预测值可能大于1 或小于0。

what离散?



what logit?

- clear
- set obs 10000
- egen p = seq(), from(0
- replace p = p/10000
- **gen lngazi = ln(p/(1-p))**
- scatter p lngazi



啥？逻辑斯蒂？

- logistics regression/ logit
- 阿耨多罗三藐三菩提——音译（鸠摩罗什）
  - 无上正等正觉
- log of it

二值选择模型的设定

- 假设个体只有两种选择，比如 $y=1$(表白)或$y=0$ (不表白)。
- 是否表白，取决于盛世美颜？毕业后的预期收入？三观一致？车子？房子？票子？……
- 假设这些解释变量都包括在向量 $\boldsymbol{x}$ 中。

两点分布

- 在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率：

$$\begin{cases} \mathrm{P}(y=1|\,\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) \\ \mathrm{P}(y=0|\,\boldsymbol{x}) = 1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases}$$

- 选择连接函数 $F(\boldsymbol{x},\boldsymbol{\beta})$为某随机变量的累积分布函数(cdf)，可保证$y$的预测值介于 [0, 1]

Probit

- 如果连接函数 $F(\boldsymbol{x}, \boldsymbol{\beta})$ 为标准正态的累积分布函数，则

$$\mathrm{P}(y=1|\,\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \int_{-\infty}^{\boldsymbol{x}} \phi(t)\,dt$$

- 此模型称为“Probit”。

Logit

- 如果连接函数 $F(\boldsymbol{x}, \boldsymbol{\beta})$ 为“逻辑分布”(logistic distribution)的累积分布函数，则

$$\mathrm{P}(y=1|\,\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$$

- 此模型称为“Logit”。

Probit vs. Logit，我pick L小姐姐

- logit=probit*1.814左右
- 更像T分布
  - 逻辑分布的密度函数关于原点对称，期望为0，方差为 $\pi^2/3$(大于标准正态的方差)。与标准正态相比，逻辑分布具有厚尾(fat tails)，更接近于自由度为7的t分布。
- 计算方便
  - 逻辑分布的cdf有解析表达式(而标准正态分布没有)，故计算Logit通常比Probit更方便。
- 更容易解释
  - Logit模型的系数估计值更易从经济上解释。

区分：几率、几率比和概率

- 对于Logit模型，记“$y=1$”的概率为p，则几率(odds)或相对风险(relative risk)为：

$$\frac{p}{1-p}=\exp(\boldsymbol{x}'\boldsymbol{\beta})$$

- 在瑞德西韦临床试验中，“$y=1$”表示“生”，“$y=0$”表示“死”。如几率为2，则存活概率是死亡概率的两倍。

对数几率

- 将上页方程两边取对数，可得“对数几率”(log-odds)：

$$\ln\left(\frac{p}{1-p}\right)=\boldsymbol{x}'\boldsymbol{\beta}$$

- β_j表示解释变量x_j增加一个微小量引起“对数几率”的边际变化。

- 或把β_j视为半弹性，即 x_j 增加一单位引起几率的变化百分比。比如，$\hat{\beta}=0.12$，意味着 x_j 增加一单位引起几率增加12%。

几率比(again)

- 比如，$\beta_j = 0.12$，则 $\exp(\beta_j) = e^{0.12} = 1.13$，故当$x_j$增加一单位时，新几率是原几率的1.13倍，或增加13%。

- Stata 称$\exp(\beta_j)$为几率比(odds ratio)。如果解释变量至少须变化一个单位(比如性别、婚否、年龄、子女个数)，则应使用 $\exp(\beta_j)$

交互效应

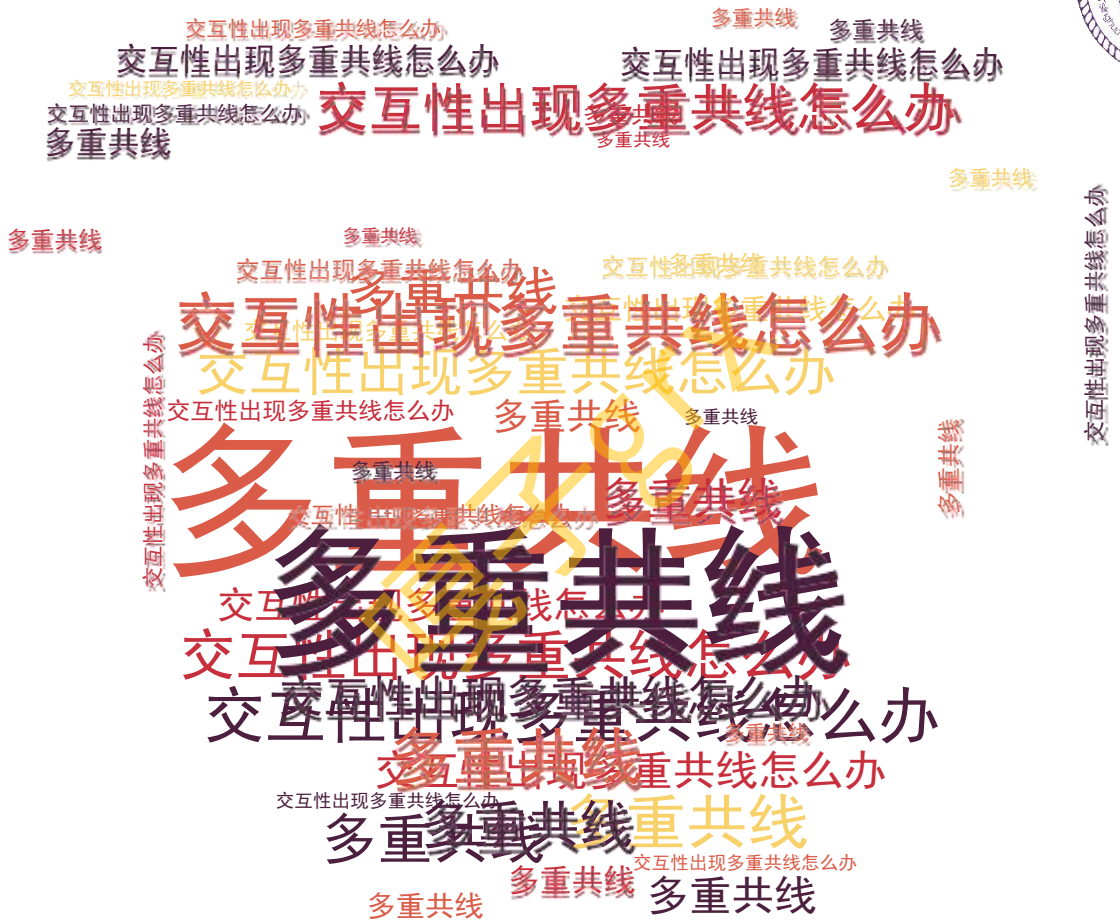
$$\mathrm{E}(y \mid \boldsymbol{x}) = \Phi(\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2) = \Phi(\cdot)$$

- 交互效应并非 $\beta_{12}\Phi'(\cdot)$，而是混合偏导数

$$\frac{\partial^2 \Phi(\cdot)}{\partial x_1 \partial x_2} = \beta_{12}\Phi'(\cdot) + (\beta_1 + \beta_{12} x_2)(\beta_2 + \beta_{12} x_1)\Phi''(\cdot)$$

- 交互效应的符号与显著性与 $\hat{\beta}_{12}$ 的符号与显著性没有必然联系

交互项的作用



交互性到底怎么用

- 连续变量*连续
  - 联动效应/协同效应/拮抗效应
- 连续*虚拟
  - 异质性
  - 分组检验
- 虚拟*虚拟 DID
- 问：调节效应、中介效应、交互效应一样吗？

新冠肺炎和流感下美国老年人决策

•研究问题：疫情当下，在美国政府无作为时，美国老年人是否购买私人医疗保险

•因变量：是否购买私人医疗保险

•自变量：linc (家庭收入的对数), hstatusg（自我评估的健康状况虚拟变量)，adl (日常生活中受限活动数目，number of limitations on activities of daily living)，chronic (慢性病数目)， age (年龄)，age2(年龄平方)，female (是否女性)，educyear (教育年限)，married (是否结婚)， hisp (是否拉丁裔)，white (是否白人)

•数据来源：美国人口普查局www.cencus.gov

内生性来源?

单击此处添加副标题

- 遗漏变量
- 选择性偏误?
- 反向因果?
- 测量误差?

教你一眼审稿就拒

- 问：加入IV后系数会变大还是变小？

清华大学学生学习与发展指导中心
乐学

清华大学公共管理学院党委研究生工作组
公管声音

谢　谢

2020春清华定量俱乐部

清华大学学生学习与发展指导中心

清华大学公共管理学院党委研究生工作组