10.24：聚类标准误

1. 标准误和标准差的区别

* 标准差 (SD)定义：

标准差 是一个描述数据分布离散程度的统计量。它衡量的是数据集中的各个观测值与均值之间的平均偏差。

标准差越大，表示数据的波动性越大；标准差越小，表示数据的波动性越小。

* 标准误 (SE)定义：

标准误 是一个描述统计量（如样本均值、回归系数等）估计值的变异性或不确定性的统计量。它衡量的是样本统计量抽样分布的标准差。

标准误越小，表示样本统计量的估计越精确；标准误越大，表示样本统计量的估计越不精确。

1. 聚类相关是什么？

（笔记本上笔记）

1. 普通稳健标准误/异方差稳健标准误/聚类稳健标准误的选择

* 异方差稳健标准误：模型中出现异方差但是观测值之间相互独立
  + vce(robust) robust r
  + 倾向于截面数据使用
* 聚类标准误：异方差，观测值组内相关
  + cluster(group)
  + cluster(id) 其中id为个体变量
  + 倾向于面板数据使用
* Newey and West 异方差自相关一致性估计(HAC)
  + 时间序列数据
  + *tsset firm\_identifier time\_identifier*

*newey dependent\_variable independent\_variables, lag(lag\_length) force*

* 一维聚类调整 异方差-序列相关稳健型标准误 (HAC)

*\**-双向固定效应模型（考虑其他类型的聚类效应）

* *xtset id year*
* *xtreg y x i.year, fe robust*
* *xtreg y x i.year, fe vce(cluster id) //与上一条命令等价*

三、一维、二位、多维聚类标准误

* 一维聚类标准误：

·截面数据，在公司层面进行聚类，以下两种写法等价

*reg y x, cluster(id)*

*reg y x, vce(cluster id)*

·面板数据，在公司层面进行聚类，以下三种写法等价

*xtset id year*

*xtreg y x, fe cluster(id)*

*xtreg y x, fe vce(cluster id)*

*xtreg y x, fe robust*

Tips： xtreg情况下使用robust会默认按照个体变量进行聚类，然后计算

*xtreg y x1 x2,fe robust*(固定效应模型)

等价于

*xtreg y x1 x2,fe cluster(id)*

* 二维聚类标准误：

如果要允许同公司或同年份的观测值具有相关性，需要从两个维度（公司和年份）进行聚类。

*cgmreg*

*vce2way*

*vcemway*

+Stata实操对比

·其他模型中的二维聚类标准误使用

logitprobit/tobit模型都可以进行二维聚类

*E.g. logit y x, regress cluster(cluster\_variable)*

1. 什么时候使用聚类标准误

1. 面板数据

在面板数据分析中，观测值通常来自不同的个体（如公司、个人、国家等），并且这些个体内部的观测值可能是相关的。例如，同一个公司在不同年份的数据之间可能存在相关性。在这种情况下，使用聚类标准误可以控制这种内部相关性。

2. 分层抽样

如果你的数据是通过分层抽样获得的，每个层级内的观测值可能会有相似的特征，并且彼此之间存在相关性。例如，在学校级别的调查中，同所学校的学生之间的成绩可能比不同学校的学生之间的成绩更为相似。此时，使用聚类标准误可以更好地反映这种相关性。

3. 群组效应

当你怀疑某些观测值属于特定的群组，并且这些群组内部的观测值可能相互关联时，使用聚类标准误是合适的。例如，在医疗研究中，同一医院的患者数据可能比不同医院的患者数据更具相关性。

4. 时间序列数据

在时间序列数据分析中，相邻时间点的数据通常具有自相关性。虽然 newey 命令可以处理这种情况，但有时你可能还需要考虑其他类型的聚类效应，例如按年度或其他时间段进行聚类。

5. 多层次模型

在多层次模型（如多级线性模型或多级逻辑回归）中，观测值通常嵌套在多个层级中（例如学生嵌套在学校中）。在这种情况下，使用聚类标准误可以控制不同层级之间的相关性。

不推荐写法（简写vce）：

*xtreg y x, fe r cluster(group)*  当个体和聚类重叠（嵌套）时有结果，但是如果个体中某一个/某几个group发生变化，会跑不出结果。反过来也可以根据是否报错来判断被聚类的group在观测区间内是否发生变动

在ols中，*reg y x, r cluster(ind)* 会和直接使用 *reg y x, vce(cluster id)*一样

1. Q&A
2. 小样本：

在样本量足够大的时候，标准回归估计量产生的无偏参数估计是有效的，但Stata默认的标准误差估计量没有考虑样本设计，这会导致不一致的标准误差。 在线形回归分析中， 当主要样本数量不足时，由于自由度过大，回归结果的标准误的估计值可能将会严重偏低，因此，基于小样本数据进行回归分析的可能会产生过于狭窄的置信区间，从而导致系数的 t 值出现误差，此时测试的第一类错误率将增加。因此，在对小样本数据进行回归分析时有必要进行纠偏处理。

* 统计效力（statistical power）

统计效力（Statistical Power）是指在给定的显著性水平下，统计检验正确地拒绝错误的H0的能力。换句话说，在H0错误的条件下，拒绝H0的概率。统计效力通常用1-β表示，其中β是第二类错误（Type II Error）的概率，也就是当备择假设为真时未能拒绝零假设的概率。

关键概念：

第一类错误 (α): 这是在零假设实际上为真的情况下错误地拒绝零假设的概率。

P(rejcetH0|H0)，通常设定为1%、5%、10%。

第二类错误 (β): 这是在H1实际上为真的情况下错误地接受H0的概率。

P(accept H0|H1)

统计效力 (1-β): 这是在H1实际上为真的情况下正确拒绝零假设的概率。

影响统计效力的因素

样本量 (Sample Size): 样本量越大，统计效力越高。增加样本量可以提高检测到真实效应的能力。

显著性水平 (Significance Level, α): 显著性水平越低，拒绝零假设的标准越严格，从而可能导致统计效力降低。

变异程度 (Variability): 数据中的Var越高，检测效应的难度越大，统计效力可能越低。