4.4理解集群相关

曾卉琪

2020.11.05

集群相关的定义

- 集群相关标准误差指的是在同一个集群内的干扰项是相关的，但不同集群间的干扰项是不相关的。
- 例子：同一个企业不同年份的观测点是一个集群，同一个企业不同年份观测点的干扰项很可能是相关的，因为影响同一家企业经营的干扰因素有连续性，但不同企业观测点的干扰项可能是不相关的。

集群相关的定义

- 可以将集群g的T个观测点用向量表示为：
- $\boldsymbol{Y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g$
- $\boldsymbol{Y}_g = (Y_{g1}, Y_{g2}, \dots, Y_{gT})', \boldsymbol{X}_g = (X_{g1}, X_{g2}, \dots, X_{gT})'$
- $\boldsymbol{e} = \left(\boldsymbol{e}_{\boldsymbol{g1}}, \boldsymbol{e}_{\boldsymbol{g2}}, \dots, \boldsymbol{e}_{\boldsymbol{gT}}\right)'$
- 集群g内的干扰项结构为：
- $E\left(\boldsymbol{e}_g\boldsymbol{e}_g'\right) = \boldsymbol{\Omega}_g = \begin{bmatrix} \sigma_{g1}^2 & \sigma_{g21}^2 & \sigma_{g31}^2 & \cdots & \sigma_{gT1}^2 \\ \sigma_{g12}^2 & \sigma_{g2}^2 & \sigma_{g32}^2 & \cdots & \sigma_{gT2}^2 \\ \sigma_{g13}^2 & \sigma_{g23}^2 & \sigma_{g33}^2 & \cdots & \sigma_{gT3}^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_{g1T}^2 & \sigma_{g2T}^2 & \sigma_{g3T}^2 & \cdots & \sigma_{gT}^2 \end{bmatrix}$

集群相关的定义

- 如果有G个集群，可以将G个集群线性相关矩阵$\boldsymbol{Y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g, g = 1,2,\ldots,G$进一步叠加，表示为：
- $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$,其中：
- $\mathbf{Y} = (\boldsymbol{Y}'_1, \boldsymbol{Y}'_2, \ldots, \boldsymbol{Y}'_G)', \boldsymbol{X} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_{\boldsymbol{G}})'$,
- $\boldsymbol{e} = (\boldsymbol{e}'_{\mathbf{1}}, \boldsymbol{e}'_{\mathbf{2}}, \ldots, \boldsymbol{e}'_{\boldsymbol{G}})'$
- 其干扰项方差矩阵结构为：
- $E(\boldsymbol{ee}') = \boldsymbol{\Omega}_{cluster} = \begin{bmatrix} \boldsymbol{\Omega}_1 & 0 & 0 & \ldots & 0 \\ 0 & \boldsymbol{\Omega}_2 & 0 & \ldots & 0 \\ 0 & 0 & \boldsymbol{\Omega}_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \boldsymbol{\Omega}_G \end{bmatrix}$

集群相关的定义

- 上式矩阵中，对角线外项为0反映了集群间干扰项不相关。对角线是集群g的方差矩阵$\mathbf{\Omega}_g$。

理解集群相关

- 例子：假设我们要估计全市中学生期末考试的平均成绩，通过在不同学校进行抽样，抽取100名学生的成绩，得到100个观测点$(Y_1, Y_2, \ldots, Y_{100})$。
- 对上述的观测点求均值，得到样本均值$\frac{1}{N}\sum_{i=1}^{N} Y_i$
- 样本均值是总体均值的无偏估计：
- $E(\bar{Y}) = E\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N}\sum_{i=1}^{N} E(Y_i) = E(Y_i)$
- 若100个学生是独立抽样，则样本均值的方差为：
- $Var(\bar{Y}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var(Y_i) = \frac{\sigma^2}{N}$

理解集群相关

- 上述方差计算的时候假设了100个观测值是独立的。
- 但是学生的成绩分布可能不是独立的。我们可以考虑两种可能的情况：
  - 若得到的观测值实际上是同一个人的成绩，那么它们的成绩完全相关，样本均值的方差为：
  - $Var(\bar{Y}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right) = \frac{1}{N^2}N^2\sigma^2 = \sigma^2$
  - 此时的方差是独立分布时的分布的N倍。

理解集群相关

- 另一种可能的情况是同一所学校所有学生的成绩可能相关，但是不同学校之间的学生的成绩不相关。
- 那么在A学生和B学生的成绩相关的情况下，B学生的成绩提供了多少新的信息呢？
- 我们可以考虑估计只有常数项的回归方程：
- $Y_i = \alpha + e_i, i = 1,2,\ldots,N, E(e_i) = 0$
- 此时$E(Y_i) = \alpha$，此时可以理解为α为全市学生成绩的平均值，e_i为个体学习成绩的差异。

理解集群相关

- 此时用OLS估计系数$\hat{\alpha}$,可以得到$\hat{\alpha} = \frac{1}{N}\sum_{i}^{100} Y_i$ 。
- 若此时样本是独立抽样，则干扰项满足同方差 $E\left(e_i^2\right) = \sigma^2, E\left(e_i e_j\right) = 0$，而估计系数的方差为 $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$，此时$\boldsymbol{X} = [\boldsymbol{1} \quad \boldsymbol{1} \quad \ldots \quad \boldsymbol{1}]'$，因此
$$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{1}'1)^{-1} = \frac{\sigma^2}{N}$$
- 因此只有常数项的回归方程在同方差的情况下，得到的结果和通常使用的样本均值估计方法所得的结果一致。

理解集群相关

- 如果同一个学校的学生成绩是相关的，我们可以把同一个学校的学生当成一个集群，则可以把回归方程写成：
- $Y_{gt} = \alpha + e_{gt}$，$g = 1,2,\ldots,G$，$t = 1,2,\ldots,T$
- 此时干扰项可以分解为两部分，$e_{gt} = c_g + v_{gt}$，其中c_g为集群因素造成的学习成绩的差异，v_{gt}为学生个人因素造成的学习成绩差异。同一个学校的学生的成绩会由于c_g而产生相关关系。

忽略集群相关造成参数估计准确度被高估的程度

- 考虑一个单变量的回归方程：
- $Y_{gt} = \beta_0 + \beta_1 X_{gt} + e_{gt}$，$g = 1,2,\ldots,G$，$t = 1,2,\ldots,T$
- $e_{gt} = c_g + v_{gt}$
- $E\left(c_g\right) = 0$，$Var\left(c_g\right) = \sigma_c^2$，$E(c_g v_{gj}) = 0$
- $E\left(v_{gt}\right) = 0$，$Var\left(v_{gt}\right) = \sigma_v^2$，$E(v_{gi} v_{gj}) = 0$

忽略集群相关造成参数估计准确度被高估的程度

- 则集群内两个观测点$t=i,t=j$的干扰项的相关系数为：
- $\rho_e=\frac{Cov(e_{gi},e_{gj})}{\sigma_{e_{gi}}\sigma_{e_{gj}}}=\frac{Var(c_g)}{Var(c_g)+Var(v_{gt})}=\frac{\sigma_c^2}{\sigma_c^2+\sigma_v^2}$
- 此时，集群g干扰项的方差结构为：
- $Var(\boldsymbol{e}_g)=E(\boldsymbol{e}_g\boldsymbol{e}_g')=\boldsymbol{\Omega}_g=$

$$(\sigma_c^2+\sigma_v^2)\begin{bmatrix} 1 & \rho_e & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & \rho_e & \dots & \rho_e \\ \rho_e & \rho_e & 1 & \dots & \rho_e \\ \dots & \dots & \dots & \dots & \dots \\ \rho_e & \rho_e & \rho_e & \cdots & 1 \end{bmatrix}$$

简单Moulton因子

- 假设每个集群的解释变量都是一样的，即$X_{gt}=X_g$。且每个集群的规模相同。
- 考虑一个组内解释变量完全相同的例子：研究班级人数对学生学习成绩的影响：
- $Score_{gi}=\alpha+\beta Size_g+e_{gi}$，$i=1,2,\ldots,N$
- 简单Moulton因子$=\frac{Var(\hat{\beta}_{cluster}^{OLS})}{Var(\hat{\beta}_{homo}^{OLS})}=1+(T-1)\rho_e$，其中T为集群规模，$\rho_e$为集群内干扰项的相关系数。

推导简单Moulton因子

- $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_G)', \boldsymbol{X}_g = \underbrace{(X_g, X_g, \dots, X_g)'}_{T\text{个}X}$

- $$Var\left(\hat{\beta}^{OLS}_{cluster}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega}_1 & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega}_G \end{bmatrix}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- $$\boldsymbol{\Omega}_g = (\sigma_c^2 + \sigma_v^2)\begin{bmatrix} 1 & \rho_e & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & \rho_e & \dots & \rho_e \\ \rho_e & \rho_e & 1 & \dots & \rho_e \\ \dots & \dots & \dots & \dots & \dots \\ \rho_e & \rho_e & \rho_e & \cdots & 1 \end{bmatrix}$$

推导简单Moulton因子

- $$\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega}_1 & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega}_G \end{bmatrix}\boldsymbol{X}$$

$$= \sum_{g=1}^{G} \boldsymbol{X}_g{}'\boldsymbol{\Omega}_g\boldsymbol{X}_g$$

$$= \sum_{g=1}^{G} (X_g, X_g, \dots, X_g)(\sigma_c^2 + \sigma_v^2)\begin{bmatrix} 1 & \rho_e & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & \rho_e & \dots & \rho_e \\ \rho_e & \rho_e & 1 & \dots & \rho_e \\ \dots & \dots & \dots & \dots & \dots \\ \rho_e & \rho_e & \rho_e & \cdots & 1 \end{bmatrix}(X_g, X_g, \dots, X_g)'$$

$$= (\sigma_c^2 + \sigma_v^2)\sum_{g=1}^{G} (TX_g{}^2 + (T-1)\rho_e X_g{}^2)$$

推导简单Moulton因子

- $Var(\hat{\beta}_{homo}^{OLS}) =$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\begin{bmatrix}\boldsymbol{\Omega} & 0 & 0 & \dots & 0\\ 0 & \boldsymbol{\Omega} & 0 & \dots & 0\\ 0 & 0 & \boldsymbol{\Omega} & \dots & 0\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega}\end{bmatrix}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

- $\boldsymbol{\Omega} = (\sigma_c^2 + \sigma_v^2)\begin{bmatrix}1 & 0 & 0 & \cdots & 0\\ 0 & 1 & 0 & \dots & 0\\ 0 & 0 & 1 & \dots & 0\\ \dots & \dots & \dots & \dots & \dots\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix}$

推导简单Moulton因子

- $$\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega} & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega} \end{bmatrix}\boldsymbol{X} = \sum_{g=1}^{G} \boldsymbol{X}_g{}'\boldsymbol{\Omega}\boldsymbol{X}_g =$$

$$\sum_{g=1}^{G}(X_g, X_g, \dots, X_g)(\sigma_c^2 + \sigma_v^2)\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}(X_g, X_g, \dots, X_g)' =$$

$$(\sigma_c^2 + \sigma_v^2)\sum_{g=1}^{G} T X_g{}^2$$

推导简单Moulton因子

- $$\frac{Var(\widehat{\beta}^{OLS}_{cluster})}{Var(\widehat{\beta}^{OLS}_{homo})} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega}_1 & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega}_G \end{bmatrix}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega} & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega} \end{bmatrix}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}$$

推导简单Moulton因子

$$
= \frac{\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega}_1 & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega}_G \end{bmatrix}\boldsymbol{X}}{\boldsymbol{X}'\begin{bmatrix} \boldsymbol{\Omega} & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega} & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{\Omega} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \boldsymbol{\Omega} \end{bmatrix}\boldsymbol{X}}
$$

$$
= \frac{(\sigma_c^2 + \sigma_v^2)\sum_{g=1}^{G}(T{X_g}^2 + (T-1)\rho_e {X_g}^2)}{(\sigma_c^2 + \sigma_v^2)\sum_{g=1}^{G} T{X_g}^2} = 1 + (T-1)\rho_e
$$

简单Moulton因子

- 简单Moulton因子$=\frac{Var(\hat{\beta}^{OLS}_{cluster})}{Var(\hat{\beta}^{OLS}_{homo})}=1+(T-1)\rho_e$，其中T为集群规模，$\rho_e$为集群内干扰项的相关系数。
- 假设$\rho_e=1$，即组内的所有干扰项是完全相关的，此时Moulton因子为T。
- 当集群内观测数T增加时，Moulton因子增加。
- 很小的集群内相关系数也能导致一个很大的Moulton因子。例如，若1000个观测点分属于10个集群，组内相关系数为0.1，此时简单Moulton因子=1+(100−1)$\times 0.1=10.9$，即忽略集群相关会导致方差低估。

广义Moulton因子

- 广义Moulton因子允许集群内解释变量不完全相同，但是彼此相关，并且允许不同集群的规模不同。
- 广义Moulton因子$=\frac{Var(\hat{\beta}^{OLS}_{cluster})}{Var(\hat{\beta}^{OLS}_{homo})}=1[\frac{V(T_g)}{\bar{T}}+\bar{T}-1]\rho_x\rho_e$
- 其中，$\rho_x=\frac{\sum g\sum j\sum i\neq j(X_{ig}-\bar{X})(X_{jg}-\bar{X})}{V(X_{gj})\sum gT_g(T_g-1)}$是组内各个解释变量之间的相关系数，$T_g$为群组g的规模，$V(T_g)$为群组规模的方差，$\bar{T}$为各个群组的平均规模。

广义Moulton因子的新增的性质

- 当各个集群规模方差$V(T_g)$较大或者解释变量X_{gt}集群内的相关系数ρ_x很大时，干扰项的集群相关方差会对标准误差造成更大的影响。
- 解释变量的集群内相关性ρ_x与干扰项的集群内相关性ρ_e一样，都会影响估计系数方差，当解释变量集群内相关系数ρ_x为0时，结果和同方差的估计值一样，此时干扰项的集群方差不影响标准误差。

处理方法

- 其一，采用GLS的方法，根据干扰项的集群方差的结构，对模型进行调整之后转换。但是很难预先知道干扰项的集群方差的结构，因此实际很少采用该方法。
- 其二，使用OLS进行估计，并且估计出集群方差下OLS估计值的标准误差，这是比较常用的方法。

处理方法

- OLS估计系数的方差为：

$$Var(\hat{\beta}^{OLS}) = (X'X)^{-1}X'E[ee']X(X'X)^{-1}$$

- 把集群方差矩阵代入，可以得到：

$$Var(\hat{\beta}^{OLS}_{cluster})$$

$$= (X'X)^{-1}X'\begin{bmatrix} \mathbf{\Omega}_1 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{\Omega}_2 & 0 & \dots & 0 \\ 0 & 0 & \mathbf{\Omega}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \mathbf{\Omega}_G \end{bmatrix} X(X'X)^{-1}$$

- 此时我们用估计值$\widehat{\Omega}_g$来代替Ω_g

处理方法

- 接上，我们用残差值$\hat{e}_{gi}^2$代替σ_{gi}^2，用$\hat{e}_{gi},\hat{e}_{gj}$来代替$\sigma_{gigj}$来计算集群稳健协方差矩阵估计值。
- 因此

$$\widehat{\boldsymbol{\Omega}}_g = \hat{\boldsymbol{e}}_g \hat{\boldsymbol{e}}_g' = \begin{bmatrix} \hat{e}_{g1}^2 & \hat{e}_{g1}\hat{e}_{g2} & \hat{e}_{g1}\hat{e}_{g3} & \cdots & \hat{e}_{g1}\hat{e}_{gT} \\ \hat{e}_{g2}\hat{e}_{g1} & \hat{e}_{g2}^2 & \hat{e}_{g2}\hat{e}_{g3} & \cdots & \hat{e}_{g2}\hat{e}_{gT} \\ \hat{e}_{g3}\hat{e}_{g1} & \hat{e}_{g3}\hat{e}_{g2} & \hat{e}_{g3}^2 & \cdots & \hat{e}_{g3}\hat{e}_{gT} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{e}_{gT}\hat{e}_{g1} & \hat{e}_{gT}\hat{e}_{g2} & \hat{e}_{gT}\hat{e}_{g3} & \cdots & \hat{e}_{gT}^2 \end{bmatrix}$$

处理方法

- $$Var\left(\hat{\beta}^{OLS}\right) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\begin{bmatrix} \widehat{\boldsymbol{\Omega}}_1 & 0 & 0 & \dots & 0 \\ 0 & \widehat{\boldsymbol{\Omega}}_2 & 0 & \dots & 0 \\ 0 & 0 & \widehat{\boldsymbol{\Omega}}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \widehat{\boldsymbol{\Omega}}_G \end{bmatrix}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}[\boldsymbol{X}_1', \boldsymbol{X}_2', \dots, \boldsymbol{X}_G']\begin{bmatrix} \widehat{\boldsymbol{\Omega}}_1 & 0 & 0 & \dots & 0 \\ 0 & \widehat{\boldsymbol{\Omega}}_2 & 0 & \dots & 0 \\ 0 & 0 & \widehat{\boldsymbol{\Omega}}_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \widehat{\boldsymbol{\Omega}}_G \end{bmatrix}\begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \dots \\ \boldsymbol{X}_G \end{bmatrix}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

处理方法

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g'\widehat{\boldsymbol{\Omega}}_g\boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\textstyle\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{e}}_g\hat{\boldsymbol{e}}_g'\boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

4.5 集群相关Stata实例

student	school	score	student	school	score
1	M	71	16	W	86
2	M	72	17	W	87
3	M	73	18	W	88
4	T	74	19	R	89
5	T	75	20	R	90
6	T	76	21	R	91
7	Q	77	22	U	92
8	Q	78	23	U	93
9	Q	79	24	U	94
10	L	80	25	S	95
11	L	81	26	S	96
12	L	82	27	S	97
13	G	83	28	A	98
14	G	84	29	A	99
15	G	85	30	A	100

集群相关方差Stata实例

- 计算集群相关系数
- loneway score school

Intraclass correlation	Asy. S.E.	[95% Conf. Interval]	
0.98798	0.00677	0.97471	1.00125

Estimated SD of school effect	9.064583
Estimated SD within school	1
Est. reliability of a school mean(evaluated at n=3.00)	0.99596

集群相关方差Stata实例

- 由前面可知，总体均值的估计等价于估计以下的回归方程：

$$Y_{gt} = \alpha + c_g + v_{gt}$$

- 由上述运行结果可知：组内相关系数为$\rho_e = 0.98798$
- 此时Moulton因子=$1 + (3 - 1) \times 0.98798 = 2.97596$
- 这就意味着假设同方差而忽略集群相关，估计系数的方差会变成真实值的1/2.97596。

集群相关方差Stata实例

- sum score

Variable	Obs	Mean	Std. Dev.	Min	Max
score	30	85.5	8.803408	71	100

- $\text{Std. Err.}(\bar{Y}) = \sqrt{Var(\hat{\alpha})} = \frac{\hat{\sigma}}{\sqrt{N}} = \frac{8.803408}{\sqrt{30}} = 1.607275$

- mean score

	Mean	Std. Err.	[95% Conf. Interval]	
score	85.5	1.607275	82.21275	88.78725

集群相关方差Stata实例

- reg score

score	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
_cons	85.5	1.607275	53.20	0.000	82.21275	88.78725

- 这三个命令默认的都是观测点的干扰项不相关，但是其实它有较大的组内相关。因此正确的标准误差应该为：

$$\text{Std. Err.}\left(\hat{\beta}_{cluster}^{OLS}\right) = \text{Std. Err.}\left(\hat{\beta}_{homo}^{OLS}\right) \times \sqrt{Moulton\ factor}$$
$$= 1.607275 \times \sqrt{2.97596} = 2.7727$$

集群相关方差Stata实例

- reg score , cluster(school)

score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
_cons	85.5	2.872281	29.77	0.000	79.00245	91.99755

- 由此可以看出，如果干扰项存在正的集群相关，而我们忽略了这个相关性，就会低估标准误差，高估了参数的准确性。

4.6集群方差运用常见问题

集群方差运用常见问题

- 由于集群方差的复杂性，到目前为止都没有选择集群的统一的标准。因为当样本数量一定时，规模更大的的集群考虑更加广泛的相关项，偏差较小，但是同时更少的集群使得方差更大，估计更加不准确。
- 一般认为在没有由于集群数量过少引发问题的情况下，尽量使用更大的集群，提高方差估计的准确性。

集群方差运用常见问题

- 如果集群内为正相关，则集群方差会比同方差大，使用集群方差后，回归系数的方差会变大。
- 如果集群内为负相关，则集群方差会比同方差小。
- 如果同时还存在异方差，此时集群方差与同方差的差异，不仅取决于集群内的相关性，还取决于异方差。

集群方差运用常见问题

- 即使在模型中加入了集群固定效应，也不一定就控制了集群相关项，不一定不需要使用集群方差。
- 例如考虑模型$Y_{it} = \alpha + \beta X_{it} + f_i + e_{it}$
- 如果企业在样本期间经历了一些意外事件，这些事件会有一些持续性，就会导致企业内的自相关，但是它们并不属于企业的固定效应。