

## Topic 5: Two-Way Cluster-Robust (TWCR) Standard Errors

by Sai Zhang

**Key points:** The validity of Two-Way Cluster-Robust (TWCR) standard errors

**Disclaimer:** This note is compiled by Sai Zhang.

### 5.1 One-Way Clustering

First, consider the case of one-way clustering. The linear model with one-way clustering

$$y_{ig} = \mathbf{x}_{ig}\boldsymbol{\beta} + u_{ig}$$

where  $i$  denotes the  $i$ th of the  $N$  individuals in the sample,  $j$  denotes the  $g$ th of the  $G$  clusters, assume that

- $\mathbb{E}[u_{ig} | \mathbf{x}_{ig}] = 0$
- error independence across clusters: for  $i \neq j$

$$\mathbb{E}[u_{ig}u_{jg'} | \mathbf{x}_{ig}, \mathbf{x}_{jg'}] = 0 \quad (5.1)$$

unless  $g = g'$ , that is, errors for individuals within the same cluster may be correlated.

Grouping observations by cluster, get

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}$$

where  $\mathbf{X}_g$  has dimension  $N_g \times K$  and  $\mathbf{y}_g$  has dimension  $N_g \times 1$ , with  $N_g$  observations in cluster  $g$ . Stacking over cluster, get the matrix form of the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

with  $\mathbf{y}, \mathbf{u}$  being  $N \times 1$  vectors,  $\mathbf{X}$  being an  $N \times K$  matrix. OLS estimator gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \left( \sum_{g=1}^G \mathbf{X}_g' \mathbf{X}_g \right)^{-1} \sum_{g=1}^G \mathbf{X}_g' \mathbf{y}_g \quad (5.2)$$

then, by CLT, we have that  $\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$  where the variance matrix of the limit normal distribution  $\boldsymbol{\Sigma}$  is

$$\left( \lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G \mathbb{E}[\mathbf{X}_g' \mathbf{X}_g] \right)^{-1} \left( \lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G \mathbb{E}[\mathbf{X}_g' \mathbf{u}_g' \mathbf{u}_g \mathbf{X}_g] \right) \times \left( \lim_{G \rightarrow \infty} \frac{1}{G} \sum_{g=1}^G \mathbb{E}[\mathbf{X}_g' \mathbf{X}_g] \right)^{-1} \quad (5.3)$$

If the primary source of clustering is due to group-level common shocks, a useful approximation is that for the  $j$ th regressor, the default OLS variance estimate based on  $s^2(\mathbf{X}'\mathbf{X})^{-1}$  should be inflated by  $\tau_j \approx 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1)$ , where

- $s$  is the estimated standard deviation of the error

- $\rho_{x_j}$  is a measure of within-cluster correlation of  $x_j$
- $\rho_u$  is the within-cluster error correlation
- $\bar{N}_g$  is the average cluster size

It's easy to see the  $\tau_j$  can be large even with small  $\rho_u$  (Kloek, 1981; Scott and Holt, 1982; Moulton, 1990). If assume the model for the cluster error variance matrices  $\mathbf{\Omega}_g = \mathbb{V}[\mathbf{u}_g | \mathbf{X}_g] = \mathbb{E}[\mathbf{u}_g \mathbf{u}_g' | \mathbf{X}_g]$ , and there is a consistent estimate  $\hat{\mathbf{\Omega}}_g$  of  $\mathbf{\Omega}_g$ , we can estimate  $\mathbb{E}[\mathbf{X}_g' \mathbf{u}_g \mathbf{u}_g' \mathbf{X}_g] = \mathbb{E}[\mathbf{X}_g' \mathbf{\Omega}_g \mathbf{X}_g]$  via GLS.

**Cluster-robust variance matrix estimate** consider

$$\hat{\mathbb{V}}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1} \left( \sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g \right) (\mathbf{X}'\mathbf{X})^{-1} \quad (5.4)$$

where  $\hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g \hat{\beta}$ . This estimate is consistent if

$$G^{-1} \sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g - G^{-1} \sum_{g=1}^G \mathbb{E}[\mathbf{X}_g' \mathbf{u}_g \mathbf{u}_g' \mathbf{X}_g] \xrightarrow{P} \mathbf{0}$$

as  $G \rightarrow \infty$ . An informal presentation of Eq.(5.4) is to rewrite the central matrix as

$$\hat{\mathbf{B}} = \sum_{g=1}^G \mathbf{X}_g' \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \mathbf{X}_g = \mathbf{X}' \begin{bmatrix} \hat{\mathbf{u}}_1 \hat{\mathbf{u}}_1' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{u}}_2 \hat{\mathbf{u}}_2' & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & & \hat{\mathbf{u}}_G \hat{\mathbf{u}}_G' \end{bmatrix} \mathbf{X} = \mathbf{X}' (\hat{\mathbf{u}} \hat{\mathbf{u}}' \otimes \mathbf{S}^G) \mathbf{X} \quad (5.5)$$

where  $\otimes$  denotes element-wise multiplication. The  $(p, q)$ th element of this matrix is

$$\sum_{i=1}^N \sum_{j=1}^N x_{ia} x_{jb} \hat{u}_i \hat{u}_j \cdot \mathbf{1}(i, j \text{ in the same cluster})$$

with  $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\beta}$ .

$\mathbf{S}^G$  is an  $N \times N$  indicator matrix with  $\mathbf{S}_{ij}^G = 1$  only if the  $i$ th and  $j$ th observation belong to the same cluster: it zeros out a large amount of  $\hat{\mathbf{u}} \hat{\mathbf{u}}'$  (asymptotically equivalently,  $\mathbf{u} \mathbf{u}'$ ), specifically, only  $\sum_{g=1}^G N_g^2$  out of  $N^2 = \left( \sum_{g=1}^G N_g \right)^2$  terms are not zero (sub-matrices on the diagonal). Asymptotically

- for fixed  $N_g$ ,  $\frac{1}{N^2} \sum_{g=1}^G N_g^2 \xrightarrow{G \rightarrow \infty} 0$
- for balanced clusters  $N_g = N/G$ ,  $\frac{1}{N^2} \sum_{g=1}^G N_g^2 = \frac{1}{G} \xrightarrow{G \rightarrow \infty} 0$

A strand of literature popularizes this method:

- Liang and Zeger (1986): in a generalized estimatin equations setting
- Arellano (1987): fixed effects estimator in linear panel models
- Hansen (2007): asymptotic theory for panel data where  $T \rightarrow \infty$  in addition to  $N \rightarrow \infty$  (or  $N_g \rightarrow \infty$  in addition to  $G \rightarrow \infty$  in the notation above).

## 5.2 Two-Way Clustering

Now, consider the case of two-way clustering,

$$y_{i,gh} = \mathbf{x}'_{i,gh} \boldsymbol{\beta} + u$$

where each observation may belong to **two** dimension of groups: group  $g \in \{1, \dots, G\}$  and  $h \in \{1, \dots, H\}$ , and for  $i \neq j$

$$\mathbb{E} [u_{i,gh} u_{j,g'h'} | \mathbf{x}_{i,gh}, \mathbf{j}, \mathbf{g}'\mathbf{h}'] = 0 \quad (5.6)$$

unless  $g = g'$  or  $h = h'$ , that is, errors for individuals within the same group (along either  $g$  or  $h$ ) may be correlated.

**Cluster-robust variance matrix estimate** extending the one-way clustering case, keep elements of  $\hat{\mathbf{u}}\hat{\mathbf{u}}'$  where the  $i$ th and  $j$ th observations share a cluster in **any** dimension, then similar to Eq.(5.5)

$$\hat{\mathbf{B}} = \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{GH} \right) \mathbf{X} \quad (5.7)$$

here  $\mathbf{S}^{GH}$  is an  $N \times N$  indicator matrix with  $S_{ij}^{GH} = 1$  only if the  $i$ th and  $j$ th observation share any cluster, the  $(p, q)$ th element of this matrix is

$$\sum_{i=1}^N \sum_{j=1}^N x_{ia} x_{jb} \hat{u}_i \hat{u}_j \cdot \mathbf{1}(i, j \text{ share any cluster})$$

$\hat{\mathbf{B}}$  can also be presented in one-way cluster-robust fashion:

$$\hat{\mathbf{B}} = \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{GH} \right) \mathbf{X} = \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^G \right) \mathbf{X} + \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^H \right) \mathbf{X} - \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G \cap H} \right) \mathbf{X} \quad (5.8)$$

where  $\mathbf{G}^{GH} = \mathbf{G}^G + \mathbf{G}^H - \mathbf{G}^{G \cap H}$ , with

- $\mathbf{G}^G$ :  $G_{ij}^G = 1$  only if the  $i$ th and  $j$ th observation belong to the same cluster  $g \in \{1, 2, \dots, G\}$
- $\mathbf{G}^H$ :  $G_{ij}^H = 1$  only if the  $i$ th and  $j$ th observation belong to the same cluster  $h \in \{1, 2, \dots, H\}$
- $\mathbf{G}^{G \cap H}$ :  $G_{ij}^{G \cap H} = 1$  only if the  $i$ th and  $j$ th observation belong to **both** the same cluster  $g \in \{1, 2, \dots, G\}$  and the same cluster  $h \in \{1, 2, \dots, H\}$

then, similar to one-way clustering case,

$$\begin{aligned} \hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^G \right) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &\quad + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^H \right) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &\quad - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \left( \hat{\mathbf{u}}\hat{\mathbf{u}}' \otimes \mathbf{S}^{G \cap H} \right) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \quad (5.9)$$

that is,

$$\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}] = \hat{\mathbf{V}}^G[\hat{\boldsymbol{\beta}}] + \hat{\mathbf{V}}^H[\hat{\boldsymbol{\beta}}] - \hat{\mathbf{V}}^{G \cap H}[\hat{\boldsymbol{\beta}}] \quad (5.10)$$

each of Eq.(5.10) can be separately computed by OLS of  $\mathbf{y}$  on  $\mathbf{X}$ , with variance matrix estimates  $\hat{\mathbf{V}}$  based on

- clustering on  $g \in \{1, 2, \dots, G\}$
- clustering on  $h \in \{1, 2, \dots, H\}$
- clustering on  $(g, h) \in \{(1, 1), \dots, (G, H)\}$

**Practical considerations** It is required to know what *ways* will be potentially important for clustering, which can be tested via checking the dimension of correlations in the errors. There are several ways to test

- estimate sample covariances of  $\mathbf{X}'\hat{\mathbf{u}}$  within dimensions, test the null that the **average** of such covariances is 0: rejecting this null is sufficient (not necessary) to reject the null of no clustering (**White, 1980**)
- for **small samples**, Eq. (5.4) is biased downwards. This is corrected (in Stata) by replacing  $\hat{\mathbf{u}}_g$  with  $\sqrt{c}\hat{\mathbf{u}}_g$ , where  $c = \frac{G}{G-1} \frac{N-1}{N-K} \simeq \frac{G}{G-1}$ . For two-way clustering (Eq. 5.8), there are 2 ways of correction:
  - choose correction terms for each of the 3 components:

$$c_1 = \frac{G}{G-1} \frac{N-1}{N-K}, c_2 = \frac{H}{H-1} \frac{N-1}{N-K}, c_3 = \frac{I}{I-1} \frac{N-1}{N-K}$$

with  $I$  being the number of unique clusters determined by  $G \cap H$

- choose a constant terms for all components:

$$c = \frac{J}{J-1} \frac{N-1}{N-K}$$

with  $J = \min(G, H)$

- $\hat{\mathbf{V}}[\hat{\boldsymbol{\beta}}]$  might have negative elements on the diagonal (Eq. 5.10), informly, this is more likely to arise when clustering is done over the same groups as the fixed effects, also
-

## References

- Manuel Arellano. Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431–434, 1987.
- Christian B Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when  $t$  is large. *Journal of Econometrics*, 141(2):597–620, 2007.
- Teunis Kloek. Ols estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica: Journal of the Econometric Society*, pages 205–207, 1981.
- Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.
- Brent R Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, pages 334–338, 1990.
- Andrew J Scott and D Holt. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American statistical Association*, 77(380):848–854, 1982.
- Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.