Cluster-robust inference

We can relax the assumption that observations are independent over $i = 1, 2, ..., n$ in various *limited* ways

One example with cross-section data occurs when the individual units of observation can be grouped in natural ways

Examples include:

- firms or plants which operate in the same industries/sectors
- households which are located in the same villages/counties/states

Survey data is often collected using stratified/clustered sampling frameworks

Suppose that each of our $i = 1, ..., n$ individual observations is allocated to one of $c = 1, ..., C$ clusters

Let y_{ic} denote the dependent variable for observation i in cluster c, with the vector of observations on the explanatory variables (x_{ic}) and the additive error term (u_{ic}) defined similarly

We have a linear model

$$y_{ic} = x'_{ic}\beta + u_{ic} \quad \text{for } i = 1, ..., n \text{ and } c = 1, ..., C$$

Let n_c denote the number of observations in cluster c, with $n_1 + n_2 + ... + n_C = \sum_{c=1}^{C} n_c = n$

Now order the observations so that the first n_1 observations are in cluster 1, the next n_2 observations are in cluster 2, and so on

Define y_c to be the $n_c \times 1$ vector containing the observations on y_{ic} for the n_c observations in cluster c. Similarly define the $n_c \times 1$ vector u_c, and the $n_c \times K$ matrix X_c. Each row of X_c contains the row vector x'_{ic} for an observation in cluster c

The model can then be written in the form

$$y_c = X_c\beta + u_c \quad \text{for} \;\; c = 1, ..., C$$

Then define the $n \times 1$ vectors y and u, and the $n \times K$ matrix X

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_C \end{pmatrix} \qquad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_C \end{pmatrix} \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_C \end{pmatrix}$$

The model can now be written as

$$y = X\beta + u$$

The OLS estimator of β is

$$\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = \left(\sum_{c=1}^{C} X_c'X_c\right)^{-1}\left(\sum_{c=1}^{C} X_c'y_c\right)$$

The observations on (y_{ic}, x'_{ic}) are now assumed to be **independent across clusters**, but not within clusters. This allows the error terms u_{ic} and u_{jc} to be correlated for different observations in the same cluster c, with general patterns of heteroskedasticity and within-cluster correlations

Asymptotic results are now derived allowing the **number of clusters** $C \to \infty$

The key assumptions that we maintain here are

$$\mathrm{E}(u_{ic}|x_{ic}) = 0$$

$$\mathrm{E}(u_{ic}u_{jd}|x_{ic}, x_{jd}) = 0 \qquad \text{for } \; i \neq j, \; \text{unless } c = d$$

These assumptions imply

$$\mathrm{E}(uu'|X) = \Omega = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_C \end{pmatrix}$$

where each Ω_c is a symmetric, but otherwise unrestricted, $n_c \times n_c$ matrix

Then for a large number of clusters, we have the limit distribution result

$$\sqrt{C}(\widehat{\beta}_{\mathrm{OLS}} - \beta) \xrightarrow{\mathrm{D}} \mathrm{N}(0, V)$$

i.e. the OLS estimator is consistent and asymptotically normal

A consistent estimator of V is provided by

$$\widehat{V} = C(X'X)^{-1}\left(\sum_{c=1}^{C} X_c'\widehat{u}_c\widehat{u}_c'X_c\right)(X'X)^{-1}$$

where $\widehat{u}_c = y_c - X_c\widehat{\beta}_{\text{OLS}}$ is the vector of OLS residuals for the n_c observations in cluster c

Hence we have the asymptotic approximation

$$\widehat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \text{N}(\beta, \widehat{V}/C)$$

The variance of $\widehat{\beta}_{\text{OLS}}$ can then be estimated using the **cluster-robust variance matrix** estimator

$$\widehat{V}/C = (X'X)^{-1}\left(\sum_{c=1}^{C} X_c'\widehat{u}_c\widehat{u}_c'X_c\right)(X'X)^{-1}$$

This allows cluster-robust standard errors, cluster-robust t-statistics and cluster-robust Wald test statistics to be calculated

Cluster-robust standard errors and test statistics can be obtained in Stata using the vce(cluster group_id) option, where group_id is a variable which takes the same value for all observations in the same cluster, e.g.

reg y x1 x2, vce(cl village)

Intuition: let $\overline{y}_c$ be the sample mean of the n_c observations on y_{ic} in cluster c, and let $\overline{X}_c$ be the $1 \times K$ vector of corresponding sample means for each of the explanatory variables

Under the stated assumptions, we could estimate β consistently (as $C \to \infty$) using the implied linear model for these cluster means

$$\overline{y}_c = \overline{X}_c\beta + \overline{u}_c \quad \text{for } c = 1, ..., C$$

in our sample of C independent observations on the cluster means $(\overline{y}_c, \overline{X}_c)$

The OLS estimator here would be asymptotically normal (as $C \to \infty$)

Using all n observations is generally more efficient (unless the observations on (y_{ic}, x'_{ic}) are perfectly correlated within each cluster)

The cluster-robust variance matrix estimator simplifies to White's (1980) heteroskedasticity-robust variance matrix estimator in the special case with only one observation in each cluster

Various small 'sample' corrections are available in this context, analogous to using $\widehat{\sigma}^2_{\text{OLS}} = \widehat{u}'\widehat{u}/(n-K)$ and the $t(n-K)$ distribution (rather than $\widehat{\sigma}^2_{\text{ML}} = \widehat{u}'\widehat{u}/n$ and the standard normal distribution)

Close connection to methods used with panel data (where repeated observations over time on the same individual may be thought of as 'clusters')

[Cameron and Trivedi, section 24.5; Wooldridge, section 20.3]