Unbiased Leave-Out Clustered Standard Errors

Raffaele Saggio

October 9, 2023

This document is written in support of the MATLAB function `KSS_SE` which permits to compute the leave-out clustered SEs introduced by Kline et al. (2020)—KSS henceforth—in a linear regression model.

1 Introduction

Consider a regression equation of the form

$$y_{ij} = x_{ij}'\beta + \varepsilon_{ij} \qquad j = 1, ..., J; \qquad i = 1, ...n_j;$$

where i indexes a particular observation which belongs to a cluster j and we have $N = \sum_j M_j$ total observations; x_{ij} is a vector of regressors of dimension $K \times 1$ and y_{ij} is the outcome of interest. The error terms, ε_{ij}, are assumed to be heteroskedastic and potentially correlated across observations belonging to the same cluster j with a block-diagonal variance-covariance matrix given by

$$\Omega = \begin{bmatrix} \mathbf{\Omega}_1 & 0 & 0 & 0 \\ 0 & \mathbf{\Omega}_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{\Omega}_J \end{bmatrix}$$

The variance of the OLS estimator of β, $\hat{\beta}$, is given by

$$\mathbb{V}[\hat{\beta}] = \left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1} \left[\sum_{j=1}^{J} \mathbf{x}_j'\Omega_j\mathbf{x}_j\right] \left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1},$$

where $\mathbf{x}_j$ is a $n_j \times K$ matrix that stacks the regressors x_{ij} for the observations belonging to cluster j.

Estimates of $\mathbb{V}[\hat{\beta}]$ in most software packages (e.g `reghdfe`) are given by

$$\tilde{\mathbb{V}}[\hat{\beta}] = d\left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1} \left[\sum_{j=1}^{J} \left(\sum_{i}^{n_j} x_{ij}\hat{e}_{ij}\right)\left(\sum_{i}^{n_j} x_{ij}\hat{e}_{ij}\right)'\right] \left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1},$$

where d is some degrees of freedom adjustment and $\hat{e}_{ij} = y_{ij} - x_{ij}\hat{\beta}$ is the OLS residual.[1]

[1] For instance, `reghdfe` sets $d = \frac{J}{J-1}\frac{N}{N-K}$.

KSS introduces an unbiased estimate of the variance of $\mathbf{\Omega}_j$ given by

$$\hat{\mathbf{\Omega}}_j = \mathbf{y}_j(\mathbf{y}_j - \mathbf{x}_j\hat{\beta}_{-j})',$$

where $\hat{\beta}_{-j}$ is the OLS estimate of β obtained after fitting (1) leaving cluster j out; $\mathbf{y}_j$ is a $n_j \times 1$ vector that stacks the outcome variable y_{ij} for observations belonging in cluster j.

Let $\hat{\eta}_{ij}$, represent the leave-cluster out residual, i.e. $\hat{\eta}_{ij} \equiv y_{ij} - x_{ij}\hat{\beta}_{-j}$. KSS shows that the following is an unbiased estimate of $\mathbb{V}[\hat{\beta}]$

$$\hat{\mathbb{V}}[\hat{\beta}] = \left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1} \left[\sum_{j=1}^{J} \left(\sum_{i}^{n_j} x_{ij}y_{ij}\right) \left(\sum_{i}^{n_j} x_{ij}\hat{\eta}_{ij}\right)'\right] \left(\sum_{i,j}^{N} x_{ij}x_{ij}'\right)^{-1},$$

The software described in this vignette, `KSS_SE`, computes $\hat{\mathbb{V}}[\hat{\beta}]$. The readme in this repository illustrates how this function works in in an example where one is interested in estimating an event-study design using a two-way fixed effects regression.

References

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.