Introduction to Robust and Clustered Standard Errors

Miguel Sarzosa

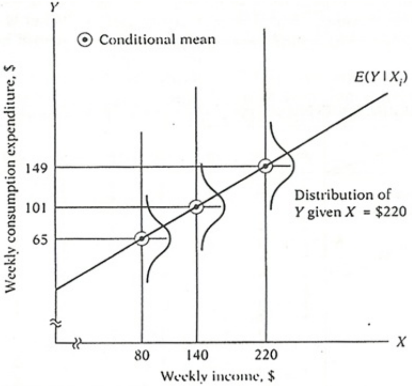
Department of Economics
University of Maryland

Econ626: Empirical Microeconomics, 2012

1. Standard Errors, why should you worry about them
2. Obtaining the Correct SE
3. Consequences
4. Now we go to Stata!

Regressions and what we estimate

- A regression does not calculate the value of a relation between two variables. In fact, it calculates a **distribution** of values of such relation. That is why, when you calculate a regression the two most important outputs you get are:
    - The conditional mean of the coefficient
    - The standard deviation of the distribution of that coefficient



Are we in the presence of a star?

- The standard errors determine how accurate is your estimation. Therefore, it affects the hypothesis testing.
- That is why the standard errors are so important: they are crucial in determining how many stars your table gets.
- And like in any business, in economics, the stars matter a lot.
- Hence, obtaining the correct SE, is critical

SE and the data

The correct SE estimation procedure is given by the underlying structure of the data.

- It is very unlikely that all observations in a data set are unrelated, but drawn from identical distributions
- For instance, the variance of income is often grater in families belonging to top deciles than among poorer families
- Some phenomena do not affect observations individually, but they affect groups of observations uniformly within each group.

SE and the data

The correct SE estimation procedure is given by the underlying structure of the data.

- It is very unlikely that all observations in a data set are unrelated, but drawn from identical distributions
- For instance, the variance of income is often grater in families belonging to top deciles than among poorer families **(heteroskedasticity)**
- Some phenomena do not affect observations individually, but they affect groups of observations uniformly within each group. **(clustered data)**

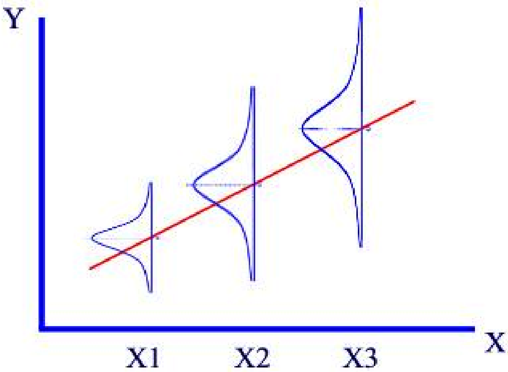
The Basic Case (i.i.d.)

$$y_i = x_i^{'}\beta + \varepsilon_i \qquad \varepsilon_i \sim \left(0, \sigma^2\right)$$

- This means that $E\left[\varepsilon\varepsilon^{'}|\mathbf{X}\right] = \Omega = \sigma^2\mathbf{I}$
- We know $\hat{\beta} = \left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}\mathbf{X}^{'}\mathbf{y} = \beta + \left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}\mathbf{X}^{'}\varepsilon$ and
- $Var\left(\beta\right) = E\left[\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}\mathbf{X}^{'}\varepsilon\varepsilon^{'}\mathbf{X}\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}\right] =$
  $\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}E\left[\mathbf{X}^{'}\varepsilon\varepsilon^{'}\mathbf{X}\right]\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1} = \sigma^2\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}$

Heteroskedasticity (i.n.i.d)

$$y_i = x_i' \beta + \varepsilon_i \qquad \varepsilon_i \sim \left(0, \sigma_i^2\right)$$



- This means that $E\left[\varepsilon\varepsilon'|\mathbf{X}\right] = \Omega = diag\left(\sigma_i^2\right)$ (big difference)

Heteroskedasticity (i.n.i.d)

- Now

$$Var(\beta) = E\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \left[\mathbf{X}'\mathbf{X}\right]^{-1}E\left[\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\right]\left[\mathbf{X}'\mathbf{X}\right]^{-1}$$

  No further simplification is possible
- Need to estimate $E\left[\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\right] = \sum_{i=1}^{N}\hat{\varepsilon}_i^2 x_i x_i'$
- Be aware that $\sum_{i=1}^{N}\hat{\varepsilon}_i^2 x_i x_i' \neq \mathbf{X}'\widehat{\varepsilon}\widehat{\varepsilon}\mathbf{X}$
- Then the Huber-Eicker-White (HEW) VC estimator is:

$$\hat{Var}\left(\hat{\beta}\right) = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\left[\sum_{i=1}^{N}\hat{\varepsilon}_i^2 x_i x_i'\right]\left[\mathbf{X}'\mathbf{X}\right]^{-1} \quad (1)$$

- Two-step procedure: 1. Get $\hat{\varepsilon}_i$, 2. Estimate (1). In Stata: `vce(robust)`

Clustered Data

- Observations are related with each other within certain groups

Example

You want to regress kids' grades on class size.
The **unobservables** of kids belonging to the same classroom will be correlated (e.g., teachers' quality, recess routines) while will not be correlated with kids in far away classrooms.

- This means, we are assuming independence across clusters but correlation within clusters. That is:

$$y_{ig} = x_{ig}^{'}\beta + \varepsilon_{ig} \qquad \text{where } g = 1, \ldots, G$$

$$E\left[\varepsilon_{ig}\varepsilon_{jg^{'}}\right] = \begin{cases} 0 & \text{if } g = g^{'} \\ \sigma_{(ij)g} & \text{if } g \neq g^{'} \end{cases} \tag{2}$$

Clustered Data

- Let us stack observations by cluster

$$\mathbf{y}_g = \mathbf{x}_g^{'}\beta + \varepsilon_g \qquad \text{where } g = 1, \ldots, G$$

- The OLS estimator of β is still $\hat{\beta} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$. (We just stacked the data)
- The variance is given by

$$Var(\beta) = E\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\Omega\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right]$$

Clustered Data

By (2), we know that

$$
\begin{aligned}
\Omega &= diag\,(\Sigma_g) \\
&= \begin{bmatrix}
\sigma^2_{(11)1} & \cdots & \sigma_{(1N_1)1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & & \vdots \\
\sigma^2_{(N_1 1)1} & & \sigma_{(N_1 N_1)1} & 0 & & 0 & & 0 & & 0 \\
0 & \cdots & 0 & \sigma^2_{(11)2} & \cdots & \sigma_{(1N_2)2} & & \vdots & & \vdots \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & & & \\
0 & \cdots & 0 & \sigma_{(N_2 1)2} & \cdots & \sigma^2_{(N_2 N_2)2} & & & & \\
 & & & & & & \ddots & \vdots & & \vdots \\
 & & & & & & & \sigma^2_{(11)g} & \cdots & \sigma_{(1N_g)g} \\
 & & & & & & & \vdots & \ddots & \vdots \\
 & & & & & & & \sigma_{(N_g 1)g} & \cdots & \sigma^2_{(N_g N_g)g}
\end{bmatrix}
\end{aligned}
$$

Clustered Data

With this in mind, we can now write the VC matrix for clustered data

$$\hat{Var}\left(\hat{\beta}\right) = \left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}\left[\sum_{i=1}^{G}\mathbf{x}_g^{'}\hat{\varepsilon}_g\hat{\varepsilon}_g^{'}\mathbf{x}_g\right]\left[\mathbf{X}^{'}\mathbf{X}\right]^{-1}$$

- In Stata: `vce(cluster` *`clustvar`*`)`. Where *`clustvar`* is a variable that identifies the groups in which onobservables are allowed to correlate.

The importance of knowing your data

- In real world you should never go with the i.i.d. case. Life is not that simple
- You need to know your data in order to choose the correct error structure and then infer the required SE calculation
- If you have aggregate variables (like class size), clustering at that level is required.
- You might think your data correlates in more than one way
  - If nested (e.g., classroom and school district), you should cluster at the highest level of aggregation
  - If not nested (e.g., time and space), you can:
    1. Include fixed-effects in one dimension and cluster in the other one.
    2. Multi-way clustering extension (see Cameron, Gelbach and Miller, 2006)

How much will my SE increase?

- Clustered SE will increase your confidence intervals because you are allowing for correlation between observations. Hence, less stars in your tables. The higher the clustering level, the larger the resulting SE.
- For one regressor the clustered SE inflate the default (i.i.d.) SE by

$$\sqrt{1+\rho_x\rho_\varepsilon\left(\bar{N}-1\right)}$$

  were ρ_x is the within-cluster correlation of the regressor, ρ_ε is the within-cluster error correlation and $\bar{N}$ is the average cluster size.
- Here it is easy to see the importance of clustering when you have aggregate regressors (i.e., $\rho_x = 1$).

Now we go to Stata!