# Microeconometrics: Clustering

Ethan Kaplan

# 1 Gauss Markov Assumptions

- OLS is minimum variance unbiased (MVUE) if
  - Linear Model: $Y_i = X_i\beta + \epsilon_i$
  - $E\left(\epsilon_i | X_i\right) = 0$
  - $V\left(\epsilon_i | X_i\right) = \sigma^2 < \infty$
  - $cov\left(\epsilon_i, \epsilon_j\right) = 0$
  - Normally distributed errors.

- What happens if we relax homoskedasticity? Uncorrelated errors?
  - Bias of $\hat{\beta}$? No!
  - Bias of $SE\left(\hat{\beta}\right)$?
    * Yes, distorted test size: OLS formula for standard errors not valid: $\sigma^2\left(X'X\right)^{-1}$

    - Up or down? Could be either (In general, positive correlation $\implies$ OLS standard errors are too low, negative correlation $\implies$ OLS standard errors are too high).
  - OLS not MVUE anymore

- This lecture will be about what to do when the homoskedasticity and uncorrelated errors assumptions are relaxed

# 2 Non-Spherical Disturbances: Examples

## 2.1 Classical OLS

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

## 2.2 Heteroskedasticity

$$\begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 \end{pmatrix}$$

## 2.3 General

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 \end{pmatrix}$$

## 2.4 General Clustered (with G clusters)

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & . & . & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & . & . & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ 0 & 0 & 0 & . & . & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ 0 & 0 & 0 & . & . & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

## 2.5 Random Effects Model

- Each cluster is structured as

$$\begin{pmatrix} \sigma^2+\sigma_G^2 & \sigma_G^2 & \sigma_G^2 \\ \sigma_G^2 & \sigma^2+\sigma_G^2 & \sigma_G^2 \\ \sigma_G^2 & \sigma_G^2 & \sigma^2+\sigma_G^2 \end{pmatrix}$$

## 2.6 Clustered AR(1) Model

$$\begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & . & . & 0 & 0 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & . & . & 0 & 0 & 0 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ 0 & 0 & 0 & . & . & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ 0 & 0 & 0 & . & . & \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

# 3 Bias in Standard Errors with Non-Spherical Disturbances

- Model Outline: Assume
  - $Y = X\beta + \epsilon$
  - $V(X) = \sigma^2_X$
  - $V(\epsilon) = \sigma^2_\epsilon$
  - $Cov\left(X_{itg}, X_{isg}\right) = \rho_x$
  - $Cov\left(X_{itg}, X_{itg'}\right) = 0$
  - $Cov\left(\epsilon_{itg}, \epsilon_{isg}\right) = \rho_\epsilon$
  - $Cov\left(\epsilon_{itg}, \epsilon_{itg'}\right) = 0$
- OLS

- $\hat{\beta}_{OLS} = (X'X)^{-1} X'Y$
- $SE\left(\hat{\beta}_{OLS}\right) = (X'X)^{-1} (X'\Omega X) (X'X)^{-1}$

- Note that
  - $X\prime X = \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^{2}$
  - $X'\epsilon = \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}\epsilon_{it}$
  - Since $X$ is one dimensional vector, we get

$$\begin{aligned} SE\left(\hat{\beta}_{OLS}\right) &= \left(X'X\right)^{-1} \left(X'\Omega X\right) \left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-2} X'\Omega X \end{aligned}$$

*

$$\Longrightarrow \quad p \lim SE\left(\hat{\beta}_{OLS}\right) = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^{2}\right)^{-2} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}\epsilon_{it}\right)^{2}$$

$$= \frac{NT\sigma_X^2\sigma_\epsilon^2 + NT(T-1)\rho_X\rho_\epsilon}{\left(NT\sigma_X^2\right)^2}$$

$$= \frac{\sigma_\epsilon^2 + (T-1)\frac{\rho_X\rho_\epsilon}{\sigma_X^2}}{NT\sigma_X^2}$$

- Implications:
  - $\rho_x > 0, \rho_\epsilon > 0 \implies$ OLS standard errors downward biased: interpretation - some of the lack of variation is not independent
  - $\rho_x > 0, \rho_\epsilon < 0 \implies$ OLS standard errors upward biased: interpretation - some of the variation is not independent

# 4 Three Types of Fixes

- Keep $\hat{\beta}$ estimate and adjust standard errors.
  - Eicker-White heteroskedasticity robust standard errors
  - Cluster-Robust standard errors (called "clustering the standard errors")
  - Use complete variance-covariance matrix for inference
- Alter the estimator of $\hat{\beta}$ in addition to using non-OLS standard errors
  - GLS - Generalized Least Squares
  - FGLS - Feasible Generalized Least Squares
  - MLE - Maximum Likelihood
- Collapse data

# 5 General Tradeoff

- By imposing structure you get greater efficiency
  - Less parameters to estimate
  - More observations per parameter
- But you could be wrong about the structure in which case you could have the wrong standard errrors

# 6 Eicker-White Heteroskedasticity Robust Standard Errors

- Heteroskedasticity robust standard errors keeps the OLS estimator but changes the standard errors by using the formula

$$V\left(\hat{\beta}_{OLS}\right) = \left(X'X\right)^{-1} X'\hat{\Omega}X \left(X'X\right)^{-1}$$

where $\hat{\Omega} =$

$$\begin{pmatrix} \hat{\epsilon}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \hat{\epsilon}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \hat{\epsilon}_3^2 & 0 & 0 \\ 0 & 0 & 0 & \hat{\epsilon}_4^2 & 0 \\ 0 & 0 & 0 & 0 & \hat{\epsilon}_5^2 \end{pmatrix}$$

- In other words:

$$V\left(\hat{\beta}_{OLS}\right) = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} \hat{\epsilon}_i^2 x_i x_i'\right) \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}$$

- Note that the sample size for estimating $\sigma_i^2$ is one so that we do not have a consistent estimate of $\sigma_i^2$.

- Tradeoff with GLS
  - Negative: Less efficient if truly heteroskedastic
  - Positive: Doesn't require knowledge of the variance-covariance matrix

# 7 Clustered Standard Errors

- When error terms are correlated within groups but not across groups and when the division of observations into groups is known, standard errors can be "clustered" or adjusted for within-group correlation.

- Clustered standard errors allow for arbitrary patterns of correlation within clusters (groups). Many clusters are needed to invoke assymptotic approximations (Donald and Lang, 2007).

## 7.1 Single Dimensional Clustering

Cluster-robust standard errors formula:

$$\left(X'X\right)^{-1} X'\hat{\Omega}X \left(X'X\right)^{-1}$$

where $\hat{\Omega} =$

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & . & . & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & . & . & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ 0 & 0 & 0 & . & . & \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ 0 & 0 & 0 & . & . & \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

In other words:

$$V\left(\hat{\beta}_{OLS}\right) = \left(\sum_{c=1}^{C} X_c' X_c\right)^{-1} \sum_{c=1}^{C} X_c' \hat{\epsilon}_c \hat{\epsilon}_c' X_c \left(\sum_{c=1}^{C} X_c' X_c\right)^{-1}$$

## 7.2 Multi-Dimensional Clustering

- Suppose correlation exists in multiple dimensions within two dimensions of groups over time (i.e. within workers over time and across workers within a certain block of time)

- Two options:
  - Choose one dimension relevant to the parameter of interest and cluster only on one dimension
  - Cluster on two dimensions

- Assumptions
  - $Y_{ijt} = X_{ijt}\beta + \epsilon_{ijt}$
  - $cov\left(\epsilon_{ijt}, \epsilon_{kjt}\right) \neq 0$
  - $cov\left(\epsilon_{ijt}, \epsilon_{imt}\right) \neq 0$
  - $cov\left(\epsilon_{ijs}, \epsilon_{ijt}\right) = 0$

- So:

$$
\begin{aligned}
V\left(\hat{\beta}_{2D}\right) &= \left(X'X\right)^{-1}\hat{Q}\left(X'X\right)^{-1} \\
\text{where } \hat{Q} &= X'\left(\hat{\Omega}S^{IJ}\right)X
\end{aligned}
$$

- $S^{IJ} = S^I + S^J - S^{I\cap J}$ where $S^K$ is the cluster matrix for dimension $K$

$$\begin{aligned}\hat{Q} &= X'\left(\hat{\epsilon}\hat{\epsilon}'S^{IJ}\right)X = \\ & X'\left(\hat{\epsilon}\hat{\epsilon}'S^{I}\right)X + X'\left(\hat{\epsilon}\hat{\epsilon}'S^{J}\right)X - \\ & X'\left(\hat{\epsilon}\hat{\epsilon}'S^{I\cap J}\right)X\end{aligned}$$

  – A cluster matrix is a matrix of zeros and ones where a zero is entered if the entry in the variance-covariance matrix is assumed to be zero and a one is entered if the entry in the variance-covariance matrix is estimated. Example: Let $S^I$ be given by (consecutive groupings):

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

- and $S^J$ be given by (odd and even groupings):

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

- Then, the intersection matrix enters a one if entries from both cluster matrices ($S^I$ and $S^J$) are one and zero otherwise:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- So $V\left(\hat{\beta}_{2D}\right) =$

$$\left(X'X\right)^{-1} X' \left(\hat{\epsilon}\hat{\epsilon}' S^I\right) X \left(X'X\right)^{-1} +$$
$$\left(X'X\right)^{-1} X' \left(\hat{\epsilon}\hat{\epsilon}' S^J\right) X \left(X'X\right)^{-1} -$$
$$\left(X'X\right)^{-1} X' \left(\hat{\epsilon}\hat{\epsilon}' S^{I\cap J}\right) X \left(X'X\right)^{-1}$$

- Thus, estimate 3 separate OLS regressions: one clustered by $S^I$, the next by $S^J$, and the third by $S^{I \cap J}$ and then compute the abve formula.

# 8 Weighted Least Squares

- We now introduce estimators where we alter the estimation of $\beta$ in addition to the standard errors. Why would we do this? Efficiency!

## 8.1 GLS

- Estimation
  - Variance-covariance matrix known: $\Omega$

- Regress $\Omega^{-\frac{1}{2}}Y = \Omega^{-\frac{1}{2}}X\beta + \Omega^{-\frac{1}{2}}\mu$
- $\hat{\beta} = (X'\Omega X)^{-1} X'\Omega Y$
- Downweights high variance observations, upweights low variance observations
- Takes into account cross-observation correlation patterns

• Positive
  - Can handle arbitrary correlation structures
  - Efficient if you know the correlation structure

• Negative
  - Relies on knowing the variance-covariance matrix $\Omega$

- Weights efficiently so doesn't estimate average treatment effect in the presence of treatment effect heterogeneity

## 8.2 FGLS

- Estimation
  - Stage 1: Run OLS - $Y = X\beta + \mu$
  - Stage 2: extract variance-covariance matrix from stage 1 - $\hat{\Omega}$ and run GLS with estimated matrix: Regress $\hat{\Omega}^{-\frac{1}{2}}Y = \hat{\Omega}^{-\frac{1}{2}}X\beta + \hat{\Omega}^{-\frac{1}{2}}\mu$
  - $\hat{\beta} = \left(X'\hat{\Omega}X\right)^{-1} X'\hat{\Omega}Y$
- Positive
  - Can handle arbitrary correlation structures

- Doesn't rely on knowing the variance-covariance matrix $\Omega$

- Negative

  - Biased in small samples: $E\left(X'\hat{\Omega}X\right)^{-1}X'\hat{\Omega}Y \neq \beta$

  - Variance-covariance matrix noisy. Note that $\hat{\beta}_{FGLS}$ is consistent for $\beta$ but $\hat{\Omega}$ is not consistent for $\Omega$. To estimate, $\hat{\Omega}$, we need to estimate $\frac{N(N+1)}{2}$ entries of the variance-covariance matrix for a sample size of $N$

  - Weights efficiently so doesn't estimate average treatment effect in the presence of treatment effect heterogeneity

# 9 Maximum Likelihood

- Can structurally model error terms - easy to allow for non-spherical disturbances
  - Note: not all distributions have an independent variance parameter - some like poisson, negative binomial, exponential have only one parameter. Others like the normal, lognormal have independent mean and variances.

- Benefits of MLE
  - Can have better small sample properties if you know the error term
  - Easier to model error structure
  - Reachers Cramer-Rao lower bound - efficient!

- Costs

- You need to know the distribution
- Not consistent if the distribution is wrong
- Can be biased in small samples even if the distribution is correct
- Doesn't generally have closed form computational formulas - have to solve simultaneously for set of first order conditions. Additional problems of knowing whether a solution to the set of first order conditions is a local/global maximum/minimum.

# 10 Structured FGLS:

## 10.1 Example - Cochrane-Orcutt

- Assume $Y_{it} = X_{it}\beta + \epsilon_{it}$ where $\epsilon_{it} = \rho\epsilon_{it-1} + \mu_{it}$

- Then follow these steps:

  1. Estimate $Y_{it} = X_{it}\beta + \epsilon_{it}$

  2. Regress $\epsilon_{it} = \rho\epsilon_{it-1} + \mu_{it}$ and obtain $\rho$

  3. Then transform data to

  $$Y_{it} - \rho Y_{it-1} = (X_{it} - \rho X_{it-1})\,\beta + \epsilon_{it}$$

  4. $\hat{\beta}$ is now correct and so are the OLS standard errors

## 10.2 Example: Newey-West

- Variance covariance matrix with each cluster assumed to equal:

$$\begin{pmatrix} \sigma^2 & \left[1-\frac{1}{M}\right]\sigma^2 & . & \left[1-\frac{K}{M}\right]\sigma^2 \\ \left[1-\frac{1}{M}\right]\sigma^2 & \sigma^2 & . & \left[1-\frac{K-1}{M}\right]\sigma^2 \\ \left[1-\frac{2}{M}\right]\sigma^2 & \left[1-\frac{1}{M}\right]\sigma^2 & . & \left[1-\frac{K-2}{M}\right]\sigma^2 \\ . & . & . & . \\ \left[1-\frac{K}{M}\right]\sigma^2 & \left[1-\frac{K-1}{M}\right]\sigma^2 & . & \sigma^2 \end{pmatrix}$$

- The above formulation, called the Newey-West estimator, allows for linear fall off in correlation of error terms within clusters

- Can be estimated using GLS or MLE

# 11 Collapsing

- Suppose that $X$ variables are the same within cluster so that

$$Y_{ig} = \alpha + \beta X_g + C_g + \epsilon_{ig}$$

- Then there is no loss in collapsing the data because there is no within cluster variation used to identify $\beta$

- Otherwise you trade off:

  – Not using variation from a correlation structure you do not know

  – Throwing away useful correlation within clusters from covariates $X$