POLSCI 630 Introduction to Regression Methods Spring 2026

Lecture 6 — Cluster Structure

Jiawei Fu, Duke University *Scribe: Jiawei Fu*

1 Overview

It is often the case that data have a clustered structure. For example, individuals may be grouped within the same classroom, school, county, or state. From a sampling perspective, clustering often arises because of cluster sampling, in which clusters or groups, rather than individual units, are drawn from a population.

This sampling scheme generally implies that outcomes for units within the same cluster are correlated through unobserved "cluster effects." Intuitively, units in the same cluster may share common environments, institutions, shocks, or social interactions that make them more similar to one another than to units in other clusters.

In previous lectures, we assumed that the data were an i.i.d. sample from a super-population. In this lecture, we relax that assumption and allow dependence within clusters while still maintaining independence across clusters.

These lecture notes draw heavily on Hansen (2022), Wooldridge (2010), and Cameron & Miller (2015).

2 Cluster

In clustered settings it is convenient to index observations as (Y_{ig}, X_{ig}), where $g = 1, \dots, G$ indexes clusters and $i = 1, \dots, n_g$ indexes individuals within the g^{th} cluster. The number of observations per cluster, n_g, may vary across clusters. The number of clusters is G, and the total number of observations is $n = \sum_{g=1}^{G} n_g$.

The linear regression model can be written at the individual level as

$$Y_{ig} = X_{ig}'\beta + e_{ig}. \tag{1}$$

It is also useful to use cluster-level notation. Let $Y_g = (Y_{1g}, ..., Y_{n_g g})'$ and $X_g = (X_{1g}, ..., X_{n_g g})'$ denote the $n_g \times 1$ vector of dependent variables and the $n_g \times k$ matrix of regressors for the g^{th} cluster. Then the linear regression model can be written in cluster notation as

$$Y_g = X_g\beta + e_g. \tag{2}$$

We can further stack the observations into full sample matrices and write the model as $Y = X\beta + e$.

The key assumption is that observations are independent across clusters but may be dependent within each cluster.

Assumption 1. *The clusters (Y_g, X_g) are mutually independent across g.*

In applications, cluster dependence is often assumed within classrooms, families, villages, firms, regions, industries, or states. The choice of clustering level is ultimately up to the researcher, but it should be justified by the context, the nature of the data, and the plausibility of dependence across observations.

3 Finite Population Results

For the linear regression model viewed as a conditional expectation function (CEF), the exogeneity assumption can be written as

$$\mathbb{E}\left[e_g|X_g\right] = \mathbb{E}\left[e_{ig}|X_g\right] = 0$$

In the clustered regression model, this requires that all relevant within-cluster interactions have been accounted for in the specification of the individual regressors X_{ig}. For example, suppose Y_{ig} measures academic performance. Then the assumption means that the achievement of any given student is unaffected by the individual controls of other students in the same school, such as their age, gender, or baseline test scores.

The OLS estimator is

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}(X'Y) \\ &= (\sum_{g=1}^{G} X_g'X_g)^{-1}(\sum_{g=1}^{G} X_g'Y_g) \\ &= (\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}X_{ig}')^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}Y_{ig})\end{aligned}$$

Under this CEF assumption, $\hat{\beta}$ is unbiased for β. Note that from (2),

$$\hat{\beta} - \beta = (\sum_{g=1}^{G} X_g'X_g)^{-1}(\sum_{g=1}^{G} X_g'e_g).$$

The proof is exactly the same as in the i.i.d. case, except that we now work with cluster-level sums:

$$\begin{aligned}\mathbb{E}\left[\hat{\beta} - \beta|X\right] &= \mathbb{E}\left[(\sum_{g=1}^{G} X_g'X_g)^{-1}(\sum_{g=1}^{G} X_g'e_g)|X\right] \\ &= (\sum_{g=1}^{G} X_g'X_g)^{-1}(\sum_{g=1}^{G} X_g'\mathbb{E}\left[e_g|X\right]) \\ &= 0\end{aligned}$$

The more interesting question is the variance. Recall that

$$Var[\hat{\beta}|X] = (X'X)^{-1}X'Var(e|X)X(X'X)^{-1}$$

Now $Var(e|X)$ is block diagonal rather than diagonal because we allow within-cluster correlation. Since observations within a cluster need not be independent, the off-diagonal entries within a

block need not be zero. By contrast, observations from different clusters are independent, so $\mathbb{E}\left[e_{ig}e_{jg'}|x_{ig}, x_{jg'}\right] = 0$ unless $g = g'$.

We denote by $\Sigma_g = \mathbb{E}\left[e_g e_g'|X_g\right]$ the $n_g \times n_g$ conditional covariance matrix of the errors within the g^{th} cluster.

$$Var[e|X] = \mathbb{E}[ee'|X] = \begin{bmatrix} \Sigma_1 & 0 & ... & 0 \\ 0 & \Sigma_2 & ... & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & ... & \Sigma_G \end{bmatrix}$$

Then, $X'Var(e|X)X = \sum_{g=1}^{G} X_g' \mathbb{E}\left[e_g e_g'|X_g\right] X_g = \sum_{g=1}^{G} X_g' \Sigma_g X_g = \Omega_n$.

Remark 1. *Another way to derive the same result is*

$$\begin{aligned} Var[\hat{\beta}|X] &= Var[(\sum_{g=1}^{G} X_g'X_g)^{-1}(\sum_{g=1}^{G} X_g'Y_g)|X] \\ &= (\sum_{g=1}^{G} X_g'X_g)^{-1} Var[\sum_{g=1}^{G} X_g'Y_g|X](\sum_{g=1}^{G} X_g'X_g)^{-1} \\ &= (X'X)^{-1} Var[\sum_{g=1}^{G} X_g'e_g|X](X'X)^{-1} \end{aligned}$$

The "meat" of the sandwich variance formula is therefore the variance of $\sum_{g=1}^{G} X_g'e_g$. Again, write $\Sigma_g = \mathbb{E}\left[e_g e_g'|X_g\right]$ for the within-cluster conditional covariance matrix. Because observations are independent across clusters, we have

$$\begin{aligned} Var[\sum_{g=1}^{G} X_g'e_g|X] &= \sum_{g=1}^{G} X_g' \mathbb{E}\left[e_g e_g'|X_g\right] X_g \\ &= \sum_{g=1}^{G} X_g' \Sigma_g X_g \\ &= \Omega_n \end{aligned}$$

We conclude that

$$V_{\hat{\beta}} = (X'X)^{-1}\Omega_n(X'X)^{-1}.$$

This differs from the formula in the independent case because it allows correlation among observations within the same cluster. Intuitively, positive within-cluster correlation reduces the amount of independent information in the sample, so the variance is typically larger than under independence.

To measure the extent of this inflation, suppose $Cor[e_{ig}, e_{jg}] = \rho_e$ for all $i \neq j$. Then a useful approximation is that, for the k^{th} regressor, the default OLS variance estimate based on $[s^2(X'X)^{-1}]_{kk}$ should be inflated by

$$1 + \rho_{x_k}\rho_e(\overline{N}_g - 1),$$

where ρ_{x_k} is a measure of the within-cluster correlation of x_{igk}, ρ_e is the within-cluster error correlation, and $\overline{N}_g$ is the average cluster size.

This important result says that the variance inflation factor increases with

1. the within-cluster correlation of the regressor

2. the within-cluster correlation of the error

3. the number of observations in each cluster.

There is effectively no clustering problem if any one of the following holds: $\rho_{x_k} = 0$, $\rho_e = 0$, or $\overline{N}_g = 1$.

4 Cluster-robust Covariance Matrix Estimator

How can we estimate this variance? If the errors were observable, then in the heteroskedastic case we would use $\mathbb{E}\left[e_i^2|X\right] = \sigma_i^2$. Similarly, with clustered dependence we have $\mathbb{E}\left[e_g e_g'|X_g\right] = \Sigma_g$, so an infeasible unbiased estimator of Ω_n is

$$\tilde{\Omega} = \sum_{g=1}^{G} X_g' e_g e_g' X_g.$$

But of course the error terms are unobserved. The natural choice is therefore to replace them with residuals:

$$\begin{aligned}
\hat{\Omega} &= \sum_{g=1}^{G} X_g' \hat{e}_g \hat{e}_g' X_g \\
&= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{l=1}^{n_g} X_{ig} X_{lg}' \hat{e}_{ig} \hat{e}_{lg} \\
&= \sum_{g=1}^{G} (\sum_{i=1}^{n_g} X_{ig} \hat{e}_{ig})(\sum_{l=1}^{n_g} X_{lg} \hat{e}_{lg})'
\end{aligned}$$

This gives the natural cluster-robust covariance estimator, CR0:

$$\hat{V}_{\hat{\beta}}^{CR0} = (X'X)^{-1} \hat{\Omega} (X'X)^{-1}.$$

This is consistent when $G \to \infty$. Note that $\hat{e}_g \hat{e}_g'$ is itself a noisy estimate of $\mathbb{E}\left[\hat{e}_g \hat{e}_g'\right]$, because there is no within-cluster averaging that would directly invoke a law of large numbers. The key point is that averaging across many clusters restores consistency.

There are also variants with better finite-sample properties.

$$\hat{V}_{\hat{\beta}} = \frac{G}{G-1} \frac{n-1}{n-K} (X'X)^{-1} \hat{\Omega} (X'X)^{-1}$$

where $\frac{G}{G-1}$ was derived by Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters G is small. The factor $\frac{n-1}{n-K}$ is analogous to the HC1 correction used in the heteroskedasticity-robust case.

4.1 Few Clusters

There are two main problems when the number of clusters is small. First, OLS tends to "overfit," so the estimated residuals are systematically too close to zero relative to the true errors. This leads to a downward-biased cluster-robust variance estimate.

Second, even after bias correction, using fitted residuals to form $\hat{\Omega}$ may still lead to over-rejection and confidence intervals that are too narrow if one relies on critical values from the standard normal distribution or even from a $t(G-1)$ distribution.

4.2 Finite-sample adjustment CR variance estimator

CR2 rescales residuals as

$$\hat{e}_g^{CR2} = (I_{n_g} - X_g(X'X)^{-1}X_g')^{-1/2}\hat{e}_g.$$

CR3 uses the stronger adjustment

$$\hat{e}_g^{CR3} = (I_{n_g} - X_g(X'X)^{-1}X_g')^{-1}\hat{e}_g.$$

As in the heteroskedasticity-robust case, CR3 is conservative in the sense that the conditional expectation of $\hat{V}_{\hat{\beta}}^{CR3}$ exceeds $V_{\hat{\beta}}$.

4.3 Bootstrap

The most important bootstrap method in this setting is the wild cluster bootstrap.

Let me explain the general idea of bootstrap. To conduct inference, we want to know the distribution of the estimator or of a test statistic. Let F_0 be the CDF of data, which is unknown. Let T_n be a statistic (like t ratio), and $G_n(t, F_0) = P(T_n \leq t)$ be the exact, finite-sample CDF of T_n. To do inference, we want to know G_n.

We have already studied asymptotic inference. It approximates $G_n(t, F_0)$ by $G_\infty(t, F_0)$, the asymptotic distribution of T_n. Many econometric statistics are asymptotically pivotal, meaning that after centering and normalization their limiting distributions do not depend on unknown population parameters. Often this limiting distribution is standard normal or chi-square. Therefore, when n is sufficiently large, $G_n(t, F_0)$ can be approximated by $G_\infty(t)$ without knowing F_0. When the sample is small, however, this approximation can be poor.

The bootstrap provides an alternative approximation to the finite-sample distribution of a statistic. It replaces the unknown distribution F_0 with an estimator F_n. One possible choice is the empirical distribution function (EDF) of the data:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

Another possibility is to specify a parametric estimator of F_0.

Regardless of the choice of F_n, the bootstrap approximation to $G_n(t, F_0)$ is $G_n(t, F_n)$. Usually, $G_n(t, F_n)$ cannot be evaluated analytically. It can, however, be approximated arbitrarily well by

Monte Carlo simulation in which random samples are drawn from F_n. The generic bootstrap procedure is as follows:

Step 1: Generate a bootstrap sample of size n by drawing from the distribution corresponding to F_n. If F_n is the EDF of the observed data, then the bootstrap sample is obtained by sampling the data with replacement.

Step 2: Compute the bootstrap statistic T_n^* from the sampled data.

Step 3: Repeat Steps 1 and 2 many times and use the empirical distribution of T_n^* to estimate probabilities, variances, confidence intervals, or critical values.

There are many versions of the bootstrap in practice.

For the few-cluster problem, the wild cluster bootstrap often performs better. The idea is to preserve the within-cluster dependence structure while perturbing the residuals at the cluster level. The resampling method works as follows. First, estimate the model under the null hypothesis H_0 that you wish to test, yielding the restricted estimate $\tilde{\beta}_{H_0}$. For example, when testing the significance of one regressor, regress y_{ig} on all components of x_{ig} except the regressor whose coefficient is set to zero under the null.

Form the restricted residual

$$\tilde{e}_{ig} = y_{ig} - x_{ig}'\tilde{\beta}_{H_0}.$$

Then obtain the b^{th} bootstrap resample as follows:

1a. Randomly assign cluster g the weight $d_g = -1$ with probability 0.5 and the weight $d_g = 1$ with probability 0.5. All observations in cluster g receive the same weight.

1b. Generate pseudo-residuals $e_{ig}^* = d_g \times \tilde{e}_{ig}$, and hence the new outcome

$$y_{ig}^* = x_{ig}'\tilde{\beta}_{H_0} + e_{ig}^*.$$

2. Re-estimate the model on the bootstrap sample and compute

$$t_b^* = \frac{\hat{\beta}_b^* - \hat{\beta}}{s_{\hat{\beta}_b^*}},$$

where $s_{\hat{\beta}_b^*}$ is the cluster-robust standard error of $\hat{\beta}_b^*$, and $\hat{\beta}$ is the OLS estimate from the original sample.

3. The bootstrap p-value is the proportion of bootstrap draws for which the bootstrap statistic is at least as extreme as the original statistic.

Again, in practice, there are many other variants.

4.4 Degrees of Freedom adjustment

With few clusters, the asymptotic distribution may be a poor approximation to the true finite-sample distribution of the statistic. As a result, critical values based on the normal approximation or even a t distribution may be inaccurate. One remedy is to adjust the test statistic so that its first two moments better match the true finite-sample distribution.

For readers interested in this issue, see Imbens & Kolesar (2016).

5 What to Cluster Over?

It is not always clear what to cluster over—that is, how to define the clusters—and there may be more than one reasonable clustering dimension.

This is not a simple question, and it has generated a substantial methodological literature.

First, suppose cluster dependence is ignored or imposed at too fine a level (for example, clustering by households instead of villages). Then the variance estimator will omit covariance terms and will generally be biased downward. Because within-cluster correlation is often positive, the resulting standard errors will be too small, leading to spurious findings of significance and overstated precision.

Second, suppose clustering is imposed at too aggregate a level (for example, clustering by states rather than villages). This does not typically create bias, but it does add many extra covariance components. As a result, the covariance estimator becomes less precise, and the reported standard errors become noisier than necessary.

5.1 Model-Based Approach

One approach is to choose the clustering level based on the model and on the likely dependence structure.

1. We should cluster broadly enough to capture the relevant dependence. Conversely, if both the regressors and the errors are plausibly uncorrelated within a potential grouping, then there is no need to cluster at that level.

2. If a key regressor is randomly assigned within clusters, or is as good as randomly assigned, then the within-cluster correlation of that regressor may be close to zero. In that case, clustering may not be necessary for inference on that regressor, even if the model errors are clustered. However, if there are other regressors of interest that are not randomly assigned within cluster, clustering may still be important for valid inference on those coefficients.

3. In some applications, dependence may arise along multiple dimensions, such as geography and time. In those cases, one may need a multi-way cluster-robust variance estimator.

5.2 Design-based Perspectives

From a design-based perspective, the appropriate clustering level depends on the sampling and treatment-assignment design.

The decision of when and how to cluster standard errors depends on how the data were sampled and how treatment was assigned, not merely on whether the outcome contains within-cluster error components.

You cluster at the level where dependence is induced by either of the following:

1. Sampling design: if units are sampled in clusters, then standard errors should reflect that clustered sampling scheme.

If one has a random sample of units from a large population with randomized treatment assignment at the unit level, there is no reason to cluster the standard errors of the least squares estimator. Doing so can be harmful, resulting in unnecessarily wide confidence intervals.

2. Treatment assignment design: cluster at the treatment-assignment level. If treatment varies only across clusters, then one should cluster at the cluster level. Do not cluster merely because residuals are correlated. If treatment is randomized at the individual level and sampling is not clustered, clustering is unnecessary and may produce overly conservative inference.

In group interaction experiments, cluster-robust standard errors are often the appropriate design-based standard errors.

References

Cameron, A. C. & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2), 317–372.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Imbens, G. W. & Kolesar, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4), 701–712.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.