

# Clustering

Greg Fischer  
MIT

April 2005

## 1 Background

You hear it all the time in seminars: “I clustered my standard errors at the village level”, “did you cluster your standard errors?” Most of us know that we can simply add the “cluster” option at that end of most STATA regression commands, but few know what exactly is going on.

## 2 When We Cluster

You’re trying to analyze the effect of a free textbook program that was randomly assigned to ten schools out of a district of twenty. You have data at the student level for each of the schools. Say there are 1,000 kids at each school. It would be great to consider your dataset as comprising 20,000 separate i.i.d observations, but the truth is kids in a school probably have a lot in common. They live in the same area so there is probably quite a bit of correlation in students’ socioeconomic backgrounds. More worrisome, unobserved variation is also likely correlated: the death of a favorite teacher or a the football winning the state championship (pretend we’re in Texas) will affect all the kids in that school.

Consider a household survey of farmers in a number of Indian villages. The quality of local roads, schools, and the like will probably not vary below the village level. There can also be neighborhood effects (“birds of a feather...”) such that eccentricities take on a local flavor. Perhaps most importantly in this setting, agricultural outcomes tend to be strongly correlated at a local level. Farmers tend to raise similar crops, are subject to the same weather, pests, and other natural hazards as well as local prices.

To the extent that these locally correlated factors are observable, we can control for them in our analysis. But problems arise when we the fact that unobserved variation is also correlated. At the statistical level, the issue is that the variance of our errors is no longer spherical and failure to account for this will lead to biased estimates of standard errors and erroneous inference. Fear not, statistical tools are available to deal with this.

### 3 The Mechanics of Clustering

If the clustering of errors in our data is ignored, OLS is no longer efficient and standard estimates of variances and covariances will be too small. Solving the first problem is a task for, in the simplest setting, GLS. Yet there are often times when we simply want to run OLS and adjust the standard errors appropriately. [Why do we do this? The only reasons I can think of are expositional simplicity (or perhaps inertia) and the fact that we may not have enough structure and information to perform FGLS, but the clustered structure goes a long way, no?]. How do we do this?

To fix ideas, consider the following model

$$y_{ig} = \mathbf{X}'_{ig}\beta + \underbrace{\alpha_g + \eta_{ig}}_{\equiv \varepsilon_{ig}}, \quad (1)$$

where  $y_{ig}$  is the dependent variable of interest for individual  $i$  in group  $g$ , the  $X$ s represent a matrix of regressors (I'll do much of the math treating  $X$  as a scalar to make the math more transparent),  $\alpha_g$  is a group specific shock that is uncorrelated across groups, and  $\eta_{ig}$  is a "nice" individual error term. The error term  $\varepsilon_{ig}$  has the following properties (which should look pretty similar to the canonical panel data assumptions where  $g$  is replaced by  $t$ ). To make matters easy, let's assume that  $X_{ig} = X_g$ , i.e. we only have group level variation in our explanatory variables, and that we have  $G$  groups with  $N$  individuals in each group.

$$\begin{aligned} E(\varepsilon_{ig}^2) &\equiv \sigma^2 = \sigma_\alpha^2 + \sigma_\eta^2 \\ E(\varepsilon_{ig}\varepsilon_{jg}) &= \sigma_\alpha^2 = \left(\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\eta^2}\right)\sigma^2 = \rho\sigma^2, \quad i \neq j \\ E(\varepsilon_{ig}\varepsilon_{jh}) &= 0, \quad g \neq h \end{aligned} \quad (2)$$

Errors within a cluster,  $g$ , exhibit a correlation of  $\rho$  (you can also think of  $\rho$  as the fraction of the variance across individuals in different clusters that is explained by the intercluster variation). Errors for individuals in different clusters are uncorrelated.

Suppose we write the regression in full matrix form  $y = \mathbf{X}\beta + \varepsilon$ . Then the variance of  $\varepsilon$  the familiar form

$$\text{Var}(\varepsilon) \equiv \mathbf{\Omega} = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma} & \vdots \\ \mathbf{0} & \dots & \mathbf{\Sigma} \end{pmatrix} = I_G \otimes \mathbf{\Sigma}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 + \sigma_\eta^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \ddots & & \vdots \\ \vdots & & \ddots & \\ \sigma_\alpha^2 & \dots & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\eta^2 \end{pmatrix} = \sigma_\eta^2 \mathbf{I}_N + \sigma_\alpha^2 i_N i_N'$$

After a little bit of work, you can show that the  $\text{Var}(\hat{\beta})$ , which equals  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  can be

reduced to

$$\text{Var}(\hat{B}) = [\sigma_\eta^2 + G\sigma_\alpha^2](\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}[1 + (G-1)\rho]. \quad (3)$$

I won't grind through the details, but the following three pieces get the result pretty quickly:

1.  $\mathbf{\Omega}$  is block diagonal with each block equal to  $\mathbf{\Sigma} = \sigma_\eta^2 \mathbf{I}_N + \sigma_\alpha^2 i_N i_N'$ .
- 2.

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{\Sigma} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\sigma_\eta^2 \mathbf{I}_N) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\sigma_\alpha^2 i_N i_N') \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\eta^2 (\mathbf{X}'\mathbf{X})^{-1} + \sigma_\alpha^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (i_N i_N') \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

3. For any column vector  $x$  of dimension  $1 \times N$  where all the elements are identical (the identical part is important. Sadly, it doesn't work in general),

$$\begin{aligned} x' i_N i_N' x &= x' \left( \sum_{i=1}^N x_i \quad \cdots \quad \sum_{i=1}^N x_i \right)' = \sum x_i \sum x_i \\ &= (nx_i)^2 = Nx'x \end{aligned}$$

which implies that

$$\mathbf{X}' (i_N i_N') \mathbf{X} = N \mathbf{X}' \mathbf{X}$$

Following the same lines as the textbook proof that  $s^2 = e'e/(n-k)$  is an unbiased estimator for  $\sigma^2$  under the Gauss-Markov assumptions, you can show that  $s^2$  will be a biased estimate under clustering (but it will be consistent if the cluster size does not increase too quickly as the sample size increases). From the results, you can show immediately that

$$\tilde{s}^2 = e'e(NG - kd)^{-1} \quad (4)$$

provides an unbiased estimate, where  $d = [1 + (N-1)\rho]$  is the *design effect* as defined by Kloeck (1981) and  $k$  is, as usual, the number of regressors.

How big of a problem is this? Moulton (1986, 1990) provides examples of the underestimate in a range of cases. The example used by Deaton (1997) is a good illustration. In an individual wage equation for the United States with only state-level explanatory variables, the design effect is more than 10. A small but significant correlation coefficient, 0.028, is combined with reasonably large cluster sizes, about 400 per state. The typical OLS measures of variance would understate errors by a factor of over three, leading to really poor inference.

Things are not quite so bad in reality. First, Scott and Holt (1982) and Pfefferman and Smith (1985), point out that this is likely to be the worst case scenario [should read these papers in more detail at some point]. When there is variance of the  $X$ s within groups or when there is variation in the cluster sizes such that  $N = \max\{N_g\}$ , this correction provides an upper bound on the variation. They also show that OLS does not produce a particularly large efficiency loss.

## 4 How do we fix it

There are two main approaches to correcting for clustering: the Moulton correction and the General Estimating Equation (GEE) framework developed by Liang and Zeger (1986), which is, as far as I can tell, just a special case of the White estimator.

### 4.1 Moulton Correction

Since we know that the true variance of  $\beta$  is bounded above by  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}[1 + (G - 1)\rho]$ , the logical next step is to follow the analogy principle and find the sample analog for this. All we're missing is  $\rho$ , so use OLS to estimate the residuals and calculate

$$\hat{\rho} = \frac{\sum_{g=1}^G \sum_{j=1}^N \sum_{i \neq j} e_{ig} e_{jg}}{GM(G-1)s^2}$$

you can then estimate the variance as

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \hat{s}^2(\mathbf{X}'\mathbf{X})^{-1}[1 + (G - 1)\hat{\rho}] \\ &= \hat{s}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Lambda}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{5}$$

where  $\mathbf{\Lambda}$  is a block diagonal matrix with one block for each cluster and where each block has ones on the diagonal and  $\hat{\rho}$  in each off diagonal position, i.e.

$$\mathbf{\Lambda} = \mathbf{I}_G \otimes \mathbf{S} \text{ and } \mathbf{S} = \begin{pmatrix} 1 & \hat{\rho} & \cdots & \hat{\rho} \\ \hat{\rho} & \ddots & \hat{\rho} & \vdots \\ \vdots & \hat{\rho} & \ddots & \hat{\rho} \\ \hat{\rho} & \cdots & \hat{\rho} & 1 \end{pmatrix}$$

I'm not quite sure why we'd want to use the second method, as the first is an easy to calculate multiple of the results that STATA or our program of choice would give us. [Confirm that these two are equivalent for multivariate regression]. Angrist and Lavy (2002) note that the Moulton correction suffers from a few drawbacks. First, the imposed error structure doesn't make sense for binary outcome variables. Second, they claim (without proof, but it's Josh so I'm inclined to trust him) that the estimates of  $\hat{\rho}$  tend to be too low. [This merits a little more understand as it doesn't quite jibe with the idea that the Moulton correction represents an upper bound].

When regressors aren't all fixed within groups, Moulton suggests the following procedure: [need to finish]

### 4.2 GEE

The Generalized Estimation equation approach is more robust alternative covariance structures. It is, in fact, just a special case of the White estimator. Calculate the OLS residuals and then form  $\mathbf{S}_g = e_g e_g'$ ,

where  $e_g$  is the vector of OLS residuals from cluster  $g$ . Then form

$$\tilde{\mathbf{\Lambda}} = \begin{pmatrix} \mathbf{S}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{S}_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{S}_G \end{pmatrix}$$

and calculate the variance using

$$\widetilde{\text{Var}}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{\Lambda}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \quad (6)$$

This is equivalent to

$$\widetilde{\text{Var}}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^G \mathbf{X}'_g e_g e'_g \mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

The nice thing about this estimator is that provided the cluster size grows slower than the sample size (in practice it's usually fixed) it provides a consistent estimator of the OLS variance-covariance matrix, that is robust to arbitrary correlation patterns within clusters and differences across clusters. Note that you can't estimate the variance of a cluster consistently with only one realization of the data, but like White and Newey-West, you can get what you need because you really care about estimating  $X'\Omega X$ , not  $\Omega$  itself.

### 4.3 Stata's Cluster Command

Here's the best news: this is what Stata calculates when you add the cluster option. Just for clarity, and because it's a rather pain to find this written down anywhere, here are the formulas for Stata's variance calculations.

- Plain old OLS (with a mild abuse of notation in the subscript of  $e_i$  that I think helps clarity):

$$V_{OLS} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2 = \frac{1}{NG} \sum_{i=1}^{NG} e_i^2.$$

- Robust (White):

$$\begin{aligned} V_{Robust} &= (\mathbf{X}'\mathbf{X})^{-1} \left( \sum_{i=1}^{NG} (e_i x_i)' (e_i x_i) \right) (\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}' e' e \mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Clustered:

$$V_{Cluster} = (\mathbf{X}'\mathbf{X})^{-1} \left( \sum_{j=1}^G u_j' u_j \right) (\mathbf{X}'\mathbf{X})^{-1}$$

where

$$u_j = \sum_{i=1}^{N_j} e_i x_i$$

and  $N_j$  is the number of observations in cluster  $j$ .

#### 4.4 Questions

- Why do Esther and Shawn Cole seem to prefer the Moulton correction to Stata's more robust command? Does it have something to do with the small sample properties of the White-like estimator? Are these any worse than those using the estimates of  $\hat{\rho}$ ? At first glance, the biases seem to be of the same magnitude.

### 5 Fixed Effects

A number of people have asked about whether fixed effects (that is including a dummy for each of the clusters) is equivalent to clustering. The short answer is no. Here's the quick intuition: suppose you have a data set that meets all of the Gauss-Markov assumptions. You then duplicate each observation  $N$  times and call each multiplicity a cluster. The clustering corrections above will return estimates numerically equal to the (correct) OLS estimates on the original data. Fixed effects, which uses only the within group estimator, can't be calculated in this case; you have no within group variation. [should probably add some more intuition as to why this is the case]

Not every case, of course, is that extreme. In general,

$$\text{Var}(\hat{\beta}_{FE}) = \sigma_\eta^2 (\mathbf{X}'\mathbf{M}_C\mathbf{X})^{-1}$$

where  $\mathbf{M}_c = \mathbf{I}_{NG} - \mathbf{C}(\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}$  (that is, the matrix that transforms  $X$  into deviations from group means) and  $\mathbf{C}$  is the matrix of fixed effects dummies, i.e.

$$\mathbf{C} = \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & a_G \end{pmatrix}, \text{ where } a_i \text{ is an } N_i \text{ vector of ones.}$$

Note that this variance does not include  $\sigma_\alpha^2$ . Once we drop the assumption that the  $\mathbf{X}$ s don't vary within

cluster, we can't simplify the true variance of the OLS estimator much beyond

$$\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Lambda}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where  $\sigma^2 = \sigma_\eta^2 + \sigma_\alpha^2$ , [or maybe we can, but I'm not sure how].

## 6 For Further Reading

- A good place to start is Deaton (1997), pp. 74-78. You'll notice that much of this write up comes straight from Deaton. While you're at it, you might as well read the entirety of Chapter 2 from this book. Yes, the title "The Analysis of Household Surveys" doesn't seem that compelling, but in my opinion the chapter ranks right up there with the Angrist and Krueger handbook chapter as a compendium of econometric tools. I also find Deaton very readable.
- The two Moulton articles are both short and worth a read. Moulton was (is?) and economist at the Bureau of Labor statistics, and his papers are very much geared to the practical. I also just came across a third article by Moulton and Randolph (1989) in *Econometrica*. I haven't read it yet, but it should probably be on the radar screen.
- Angrist and Lavy (2002) is a great paper. The only problem with it may come from the inferiority complex you'll develop when reading the discussion of empirical methods. The relevant meat ranges from pages 16 to 18. They also discuss a range of other estimators, including the two-step procedures of Baker and Fortin (2001) and Donald and Lang (2001) and Bell and McCaffrey's (2002) Biased Reduced Linearization estimator, which is similar to McKinnon and White's (1985) bias-corrected heteroskedastic-consistent covariance matrix. At some point I'll try to get a better handle on these and add them to the write-up. The Econometrics are but a tool in this paper, and the implementation is also quite interesting.
- Mark Wooldrige has on his website a nice little paper on cluster-sample methods that illuminates but doesn't resolve some of the issue that develop when the number of clusters ( $G$  in this paper and his) is small. He makes reference to the particular problems posed by diff-in-diffs estimates, with references to Bertrand, Duflo and Mullinathan's (2004) paper on the subject and the difficulties encountered in the Card and Krueger (1994) paper on the minimum wage, which attempted to exploit differences between New Jersey and Pennsylvania, really just two clusters.
- For those truly interested in econometrics, read McKinnon and White (1985), Liang and Zeger (1986), and the aforementioned Klock (1981).

## References

- Angrist, J. D. and V. Lavy (2002). The Effect of High School Matriculation Awards: Evidence from Randomized Trials. Technical report, MIT Working Paper.
- Baker, M. and N. M. Fortin (2001). Occupational gender composition and wages in Canada, 1987-1988. *Canadian Journal of Economics* 34, 345–76.
- Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology* 28.
- Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconomic Approach*. Baltimore: Johns Hopkins University Press.
- Donald, S. and K. Lang (2001, March). Inference with differences-in-differences and other panel data. Technical report, Boston University Department of Economics.
- Kloek, T. (1981, Jan.). OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated. *Econometrica* 49(1), 205–207.
- Liang, K. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- MacKinnon, J. and H. White (1985). Some heteroscedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Econometrics* 28, 205–325.
- Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics* 32, 385–397.
- Moulton, B. R. (1990, May). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics* 72(2), 334–338.
- Moulton, B. R. and W. Randolph (1989). Alternative tests of the error component model. *Econometrica* 57(3), 685–93.
- Pfefferman, D. and T. Smith (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review* 53, 37–59.
- Scott, A. and D. Holt (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 77, 848–854.