

## Clustering in the Linear Model

### 1 Introduction

This handout extends the handout on “The Multiple Linear Regression model” and refers to its definitions and assumptions in section 2. It relaxes the homoscedasticity assumption (*OLS5a*) and allows the error terms to be heteroscedastic and correlated within groups or so-called clusters. It shows in what situations the parameters of the linear model can be consistently estimated by OLS and how the standard errors need to be corrected.

The canonical example (Moulton 1986, 1990) for clustering is a regression of individual outcomes (e.g. wages) on explanatory variables of which some are observed on a more aggregate level (e.g. employment growth on the state level).

Clustering also arises when the sampling mechanism first draws a random sample of groups (e.g. schools, households, towns) and then surveys all (or a random sample of) observations within that group. Stratified sampling, where some observations are intentionally under- or oversampled asks for more sophisticated techniques.

### 2 The Econometric Model

Consider the multiple linear regression model

$$y_{ig} = x'_{ig}\beta + u_{ig}$$

where observations belong to a cluster  $g = 1, \dots, G$  and observations are indexed by  $i = 1, \dots, N_g$  within their cluster.  $N_g$  is the number of observations in cluster  $g$ ,  $N = \sum_g N_g$  is the total number of observations,  $y_{ig}$  is the dependent variable,  $x'_{ig}$  is a  $(K + 1)$ -dimensional row vector of  $K$

explanatory variables plus a constant,  $\beta$  is a  $(K + 1)$ -dimensional column vector of parameters, and  $u_{ig}$  is the error term.

Stacking observations within a cluster, we can write

$$y_g = X_g\beta + u_g$$

where  $y_g$  is a  $N_g \times 1$  vector,  $X_g$  is a  $N_g \times (K + 1)$  matrix and  $u_g$  is a  $N_g \times 1$  vector. Stacking observations cluster by cluster, we can write

$$y = X\beta + u$$

where  $y = [y'_1 \dots y'_G]'$  is  $N \times 1$ ,  $X$  is  $N \times (K + 1)$  and  $u$  is  $N \times 1$ .

The data generation process (dgp) is fully described by:

*CL1: Linearity*

$$y_{ig} = x'_{ig}\beta + u_{ig} \text{ and } E(u_{ig}) = 0$$

*CL2: Independence*

$$(X_g, y_g)_{g=1}^G \text{ i.i.d. (independent and identically distributed)}$$

*CL2* assumes mainly that the observations in one cluster are independent from the observations in all other clusters.

*CL3: Strict Exogeneity*

- a)  $u_{ig}|X_g \sim N(0, \sigma_{ig}^2)$
- b)  $u_{ig} \perp X_g$  (independent)
- c)  $E(u_{ig}|X_g) = 0$  (mean independent)
- d)  $Cov(X_g, u_{ig}) = 0$  (uncorrelated)

*CL3* assumes that the error term  $u_{ig}$  is unrelated to the explanatory variables ( $X_g$ ) of all observations within its cluster.

*CL4: Identifiability*

$$\text{rank}(X) = K + 1 < N$$

*CL4* assumes that the regressors are not perfectly collinear.

*CL5: Clustered Errors*

$$V(u_g|X_g) = \sigma^2\Omega_g = \sigma^2\Omega(X_g) \text{ is p.d. and finite (condit. clustering)}$$

*CL5* means that the error terms are allowed to be correlated within clusters and to have different variances conditional on  $X_g$ .

*CL6: Variance of explanatory variables*

$$E(X_g'X_g) = Q \text{ is positive definite and finite}$$

The variance-covariance of the vector of error terms in the whole sample is under *CL2* and *CL5*

$$\begin{aligned} V(u|X) &= E(uu'|X) = \sigma^2\Omega \\ &= \begin{pmatrix} \sigma^2\Omega_1 & 0 & \cdots & 0 \\ 0 & \sigma^2\Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2\Omega_G \end{pmatrix} \end{aligned}$$

where the typical diagonal element

$$\sigma^2\Omega_g = V(u_g|X_g) = E(u_gu_g'|X_g) = \sigma^2\Omega(X_g)$$

$$= \begin{pmatrix} \sigma_{1g}^2 & \rho_{12}\sigma_{1g}\sigma_{2g} & \cdots & \rho_{1N_{1g}}\sigma_{1g}\sigma_{N_{1g}} \\ \rho_{12}\sigma_{1g}\sigma_{2g} & \sigma_{2g}^2 & \cdots & \rho_{2N_{1g}}\sigma_{2g}\sigma_{N_{1g}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N_{1g}}\sigma_{1g}\sigma_{N_{1g}} & \rho_{2N_{1g}}\sigma_{2g}\sigma_{N_{1g}} & \cdots & \sigma_{N_{1g}}^2 \end{pmatrix}$$

is the variance covariance matrix of the error terms within cluster  $g$  and all its elements are a function of  $X_g$ . The decomposition into  $\sigma^2$  and  $\Omega_g$  is arbitrary but useful.

### 3 A Generic Case: Cluster Specific Random Effects

Suppose as Moulton(1986) that the error term  $u_{ig}$  consists of a cluster specific random effect  $\alpha_g$  and an individual effect  $\nu_{ig}$

$$u_{ig} = \alpha_g + \nu_{ig}$$

Assume that the individual error term is homoscedastic and independent across all observations

$$V(\nu_{ig}|X_g) = \sigma_\nu^2$$

$$Cov(\nu_{ig}, \nu_{jg}|X_g) = 0, i \neq j$$

and that the cluster specific effect is homoscedastic and uncorrelated with the individual effect

$$V(\alpha_g|X_g) = \sigma_\alpha^2$$

$$Cov(\alpha_g, \nu_{ig}|X_g) = 0$$

The cluster specific effect  $\alpha_g$  is under OLS3 at least uncorrelated with  $X_g$  and can therefore be treated as a *random effect*:

$$Cov(\alpha_g, X_g) = 0.$$

The resulting variance-covariance structure within each cluster  $g$  is then

$$V(u_g|X_g) = \sigma^2\Omega_g = \sigma^2[\rho\iota\iota' + (1-\rho)I] = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

where  $\sigma^2 = \sigma_\alpha^2 + \sigma_\nu^2$ ,  $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\nu^2)$  and  $\iota$  is a  $(N_g \times 1)$  column vector of ones. In a less restrictive version,  $\sigma_g^2$  and  $\rho_g$  are allowed to be cluster specific as a function of  $X_g$ . We call this structure *equicorrelated* errors.

Note: this structure is identical to a panel data random effects model with many individuals  $g$  observed over few time periods  $i$ .

## 4 Estimation with OLS

The parameter  $\beta$  can be estimated with OLS as

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'y$$

The OLS estimator of  $\beta$  remains unbiased in small samples under *CL1*, *CL2*, *CL3c*, *CL4*, *CL5* and *CL6* and normally distributed additionally assuming *CL3a*. It is consistent and approximately normally distributed under *CL1*, *CL2*, *CL3d*, *CL4*, *CL5* and *CL6b* in samples with a large number of clusters. However, the OLS estimator is not efficient any more. More importantly, the usual standard errors of the OLS estimator and tests (*t*-, *F*-, *z*-, Wald-) based on them are not valid any more.

## 5 Estimating Correct Standard Errors

The small sample covariance matrix of  $\hat{\beta}_{OLS}$  is under *CL3c* and *CL5*

$$V(\hat{\beta}_{OLS}|X) = \sigma^2 (X'X)^{-1} [X'\Omega X] (X'X)^{-1}$$

and differs from usual OLS where  $V(\hat{\beta}_{OLS}|X) = \sigma^2(X'X)^{-1}$ . Consequently, the usual estimator  $\hat{V}(\hat{\beta}_{OLS}|X) = \hat{\sigma}^2(X'X)^{-1}$  is incorrect. Usual small sample test procedures, such as the *F*- or *t*-Test, based on the usual estimator are therefore not valid.

With the number of clusters  $G \rightarrow \infty$  and fixed cluster size  $N_g = N/G$ , the OLS estimator is asymptotically normally distributed under *CL1*, *CL2*, *CL3d*, *CL4*, *CL5* and *CL6*

$$\sqrt{G}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$$

where  $\Sigma = Q_{XX}^{-1} E(X'_g u_g u'_g X_g) Q_{XX}^{-1}$  and  $Q_{XX} = E(X'_g X_g)$ . The OLS estimator is therefore approximately normally distributed in samples with a large number of clusters

$$\hat{\beta} \overset{A}{\sim} N(\beta, Avar(\hat{\beta}))$$

where  $Avar(\hat{\beta}) = G^{-1}\Sigma$  can be estimated as

$$\widehat{Avar}(\hat{\beta}) = (X'X)^{-1} \left[ \sum_{g=1}^G X'_g \hat{u}_g \hat{u}'_g X_g \right] (X'X)^{-1}$$

with  $\hat{u}_g = y_g - X_g \hat{\beta}_{OLS}$  under some additional assumptions on higher order moments of  $X_g$ .

This so-called *cluster-robust* covariance matrix estimator is a generalization of Huber(1967) and White(1980).<sup>1</sup> It does not impose any restrictions on the form of both heteroscedasticity and correlation within clusters (though we assumed independence of the error terms across clusters). We can perform the usual *z*- and Wald-test for large samples using the cluster-robust covariance estimator.

Note: the cluster-robust covariance matrix is consistent when the number of clusters  $G \rightarrow \infty$ . In practice we should have 50 or more clusters.

Bootstrapping is an alternative method to estimate a cluster-robust covariance matrix under the same assumptions. See the handout on “The Bootstrap”. Clustering is addressed in the bootstrap by randomly drawing clusters  $g$  (rather than individual observations  $ig$ ) and taking all  $N_g$  observations for each drawn cluster. This so-called *block bootstrap* preserves all within cluster correlation.

## 6 Efficient Estimation with GLS

In some cases, for example with cluster specific random effects, we can estimate  $\beta$  efficiently using feasible GLS (see the handout on “Heteroscedasticity in the Linear Model” and the handout on “Panel Data”). In practice, we can rarely rule out additional serial correlation beyond the one induced by the random effect. It is therefore advisable to always use cluster-robust standard errors in combination with FGLS estimation of the random effects model.

<sup>1</sup>Note: the cluster-robust estimator is not clearly attributed to a specific author. See e.g. [http://www.stata.com/support/faqs/stat/robust\\_ref.html](http://www.stata.com/support/faqs/stat/robust_ref.html)

## 7 Special Cases

### Case 1: cluster random effects and aggregate regressors

Assume cluster specific random effects, regressors which are constant within clusters and clusters of equal size  $N_g = N/G$ :

$$y_{ig} = \beta_0 + \beta_1 \bar{x}_{g1} + \dots + \beta_K \bar{x}_{gK} + u_{ig}$$

where  $u_{ig} = \alpha_g + \nu_{ig}$  with  $\sigma^2 = \sigma_\alpha^2 + \sigma_\nu^2$ ,  $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\nu^2)$ . Then the (cluster-robust) asymptotic variance can now be estimated as

$$\widehat{Avar}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1} [1 + (N_g - 1)\widehat{\rho}]$$

where  $\widehat{\sigma}^2 (X'X)^{-1}$  is the usual OLS variance estimator,  $[1 + (N_g - 1)\rho] > 1$  is called the *Moulton factor*,  $\widehat{\sigma}^2$  and  $\widehat{\rho}$  are consistent estimators of  $\sigma^2$  and  $\rho$  respectively.

The square root of the Moulton factor measures how much the usual OLS standard errors understate the correct standard errors. For example, with cluster size  $N_g = 500$  and intracluster correlation  $\rho = 0.1$ , the correct standard errors are 7.13 times the usual OLS ones. The Moulton factor is highest when the observations within clusters are perfect clones ( $\rho = 1$ ).

### Case 2: i.i.d. errors and aggregate regressors

Assume  $u_{ig}$  i.i.d. within and across clusters then

$$\widehat{Avar}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1}$$

which is the usual OLS estimator.

### Case 3: cluster random effects in bivariate regression

Assume cluster specific random effects in a regression with one variable:

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig}$$

where  $u_{ig} = \alpha_g + \nu_{ig}$  with  $\sigma^2 = \sigma_\alpha^2 + \sigma_\nu^2$ ,  $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\nu^2)$ . Then the (cluster-robust) asymptotic variance can be estimated as

$$\widehat{Avar}(\widehat{\beta}_1) = \widehat{\sigma}^2 [(X'X)^{-1}]_{11} [1 + (N_g - 1)\widehat{\rho}_x \widehat{\rho}_u]$$

where  $[\cdot]_{11}$  means the element in the 2nd row and 2nd column,  $\rho_x$  is the within cluster correlation of  $x$  and  $[1 + (N_g - 1)\rho_x \rho] > 1$  is again called the *Moulton factor*.  $\widehat{\sigma}^2$ ,  $\widehat{\rho}_u$  and  $\widehat{\rho}_x$  are consistent estimators of  $\sigma^2$ ,  $\rho_u$  and  $\rho_x$ , respectively.

The square root of the Moulton factor measures how much the usual OLS standard errors understate the correct standard errors. For example, with cluster size  $N_g = 500$  and intracluster correlations  $\rho_u = 0.1$  and  $\rho_x = 0.1$ , the correct standard errors are 2.45 times the usual OLS ones.

### Case 4: cluster random effects and i.i.d. regressors

Assume cluster specific random effects, regressors  $x'_{ig}$  which are i.i.d. within and across clusters and clusters of equal size  $N_g = N/G$ , then

$$\widehat{Avar}(\widehat{\beta}) = \widehat{\sigma}^2 (X'X)^{-1}$$

which is the usual OLS estimator.

### Lessons from the 4 cases

Cases 1 to 4 with cluster random effects are most relevant for the analysis of cross-section data. They teach us:

- If the variable of interest is an aggregate variable, we need to correct the standard errors.
- If only control variables are aggregated, we better include cluster fixed effects (i.e. dummy variables for the groups) to take care of the random effect.
- If only control variables have an important cluster-specific component, it is better to include cluster fixed effects.
- If the variable of interest is not aggregated and has only a small cluster specific component (i.e. a lot of within-cluster variation and very little between-cluster variation), it is better to include cluster fixed effects.
- If the variable of interest is not aggregated but has an important cluster specific component (i.e. a lot of between-cluster variation and very little within-cluster variation), then including cluster fixed effects may destroy the most valuable information and we may not want to include cluster fixed effects. However, we need to correct the standard errors.

Standard errors are most easily corrected using the (more general) cluster-robust covariance from section 5. Note again, that we should have  $G = 50$  clusters or more to justify the asymptotic approximation.

In the context of panel and time series data, serial correlation beyond the ones from a random effect becomes very important. See the handout on “Panel Data: Fixed and Random Effects”. In this case standard errors need to be corrected even when including fixed effects.

## 8 Implementation in Stata 10.0

Stata reports the cluster-robust covariance estimator with the `vce(cluster)` option, e.g.<sup>2</sup>

```
webuse auto7.dta
regress price weight, vce(cluster manufacturer)
matrix list e(V)
```

Note: Stata multiplies  $\widehat{V}$  with  $(N - 1)/(N - K) \cdot G/(G - 1)$  to correct for degrees of freedom in small samples. Stata reports  $t$ - and  $F$ -statistics with  $G - 1$  degrees of freedom.

We can also estimate a heteroscedasticity robust covariance using a nonparametric block bootstrap. For example with either of the following,

```
regress price weight, vce(bootstrap, reps(100) cluster(manufacturer))
bootstrap, reps(100) cluster(manufacturer): regress price weight
```

The cluster specific random effects model is efficiently estimated by FGLS. For example,

```
xtset manufacturer_grp
xtreg price weight, re
```

In addition, cluster-robust standard errors are reported with

```
xtreg price weight, re vce(cluster manufacturer)
```

<sup>2</sup>There are only 23 clusters in this example dataset used by the Stata manual. This is not enough to justify using large sample approximations.

## References

### Advanced textbooks

- Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press. Sections 24.5.
- Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press. Sections 7.8 and 11.54.

### Companion textbooks

- Angrist, Joshua D. and Jörn-Steffen Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press. Chapter 8.

### Articles

- Huber, P. J. (1967), The behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA: University of California Press, 1, 221-238.
- Kloek T. (1981), OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated, *Econometrica*, 49(1), 205-207.
- Liang, Kung-Yee And Scott L. Zeger (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, 73(1), 13-22.
- Moulton, B. R. (1986), Random Group Effects and the Precision of Regression Estimates, *Journal of Econometrics*, 32(3), 385-397.
- Moulton, B. R. (1990), An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units, *The Review of Economics and Statistics*, 72, 334-338.
- White, H. (1980), A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica* 48, 817-838.