Kurt Schmidheiny
University of Basel

Clustering in the Linear Model

matrix-free

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 (*OLS4a*) 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 than 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

$$y_{gi} = \beta_0 + \beta_1 x_{gi1} + ... + \beta_K x_{giK} + u_{gi}$$

where observations belong to a cluster $g = 1, ..., G$ and observations are indexed by $i = 1, ..., M$ within their cluster. G is the number of clusters,

M is the number of observations per cluster, and $N = \sum_g M = GM$ is the total number of observations. For notational simplicity, M is assumed constant in this handout. It is easily generalized to a cluster specific number M_g. y_{gi} is the dependent variable, $x_{gi1}, ..., x_{giK}$ are K explanatory variables, $\beta_0, ..., \beta_K$ are $K + 1$ parameters, and u_{gi} is the error term.

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

CL1: Linearity

$y_{gi} = \beta_0 + \beta_1 x_{gi1} + ... + \beta_K x_{giK} + u_{gi}$ and $E[u_{gi}] = 0$

CL2: Independence

$\{x_{g11}, ..., x_{gMK}, y_{g1}, ..., y_{gM}\}_{g=1}^{G}$
i.i.d. (independent and identically distributed)

CL2 assumes that the observations in one cluster are independent from the observations in all other clusters. It does *not* assume independence of the observations *within* clusters.

CL3: Strict Exogeneity

a) $u_{gi}|x_{g11}, ..., x_{gMK} \sim N(0, \sigma_{gi}^2)$

b) $\forall j, k : u_{gi} \perp x_{gjk}$ (independent)

c) $E[u_{gi}|x_{g11}, ..., x_{gMK}] = 0$ (mean independent)

d) $\forall k, j : Cov[x_{gjk}, u_{gi}] = 0$ (uncorrelated)

CL3 assumes that the error term u_{gi} is unrelated to all explanatory variables of all observations within its cluster.

CL4: Clustered Errors

$V[u_{gi}|x_{g11}, ..., x_{gMK}] = \sigma_{gi}^2 > 0$ and $< \infty$
$Cov[u_{gi}, u_{gj}|x_{g11}, ..., x_{gMK}] = \rho_{gij}\sigma_{gi}\sigma_{gj} < \infty$, for all $i \neq j$

CL4 means that the error terms are allowed to have different variances and to be correlated within clusters conditional on all explanatory variables

of all observations within the cluster.

Under *CL2*, *CL3c* and *CL4*, the conditional variances and covariances across all error terms are

$$V(u_{gi}|x_{g11},...,x_{gMK}) = \sigma_{gi}^2$$

$$Cov(u_{gi},u_{gj}|x_{g11},...,x_{gMK}) = \rho_{gij}\sigma_{gi}\sigma_{gj}, i \neq j$$

$$Cov(u_{gi},u_{hj}|x_{g11},...,x_{gMK},x_{h11},...,x_{hMK}) = 0, i \neq j, g \neq h$$

CL5: Identifiability

$(1, x_{gi1}, \cdots, x_{giK})$ are not linearly dependent
$0 < V[x_{gik}] < \infty$ and $0 < \widehat{V}[x_{gik}]$

CL5 assumes that the regressors have identifying variation (non-zero variance) and are not perfectly collinear.

3 A Special Case: Random Cluster-specific Effects

Suppose as Moulton(1986) that the error term u_{gi} consists of a cluster specific random effect c_g and an individual effect v_{gi}

$$u_{gi} = c_g + v_{gi}$$

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

$$E[v_{gi}|x_{g11},...,x_{gMK}] = 0$$

$$V[v_{gi}|x_{g11},...,x_{gMK}] = \sigma_v^2$$

$$Cov[v_{gi},v_{gj}|x_{g11},...,x_{gMK}] = 0,\ i \neq j$$

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

$$E[c_g|x_{g11},...,x_{gMK}] = 0$$

$$V[c_g|x_{g11},...,x_{gMK}] = \sigma_c^2$$

$$Cov[c_g,v_{gi}|x_{g11},...,x_{gMK}] = 0$$

The resulting variances and covariances of the combined error term $u_{gi} = c_g + v_{gi}$ are then within each cluster g

$$V[u_{gi}|x_{g11},...,x_{gMK}] = \sigma_u^2$$

$$Cov[u_{gi},u_{gj}|x_{g11},...,x_{gMK}] = \rho_u\sigma_u^2, i \neq j$$

where $\sigma_u^2 = \sigma_c^2 + \sigma_v^2$ and $\rho_u = \sigma_c^2/(\sigma_c^2 + \sigma_v^2)$. This structure is called *equicorrelated* errors. In a less restrictive version, σ_u^2 and ρ_u are allowed to be cluster specific as a function of $x_{g11},...,x_{gMK}$.

Note: this structure is formally identical to a random effects model for panel data with many "individuals" g observed over few "time periods" i. The cluster specific random effect is also called an *unrelated effect*.

4 Estimation with OLS

The parameter β can be estimated with OLS by regressing y_{gi} on a constant and on $x_{gi1}, \cdots, x_{giK}$. In the special case with one regressor x_{gi}, the resulting OLS estimators of β_0 and β_1 are:

$$\widehat{\beta}_1 = \frac{\sum_{g=1}^{G}\sum_{i=1}^{M}(x_{gi}-\bar{x})(y_{gi}-\bar{y})}{\sum_{g=1}^{G}\sum_{i=1}^{M}(x_{gi}-\bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

where $\bar{y} = 1/GM\sum_g\sum_i y_{gi}$ and $\bar{x} = 1/GM\sum_g\sum_i x_{gi}$.

The OLS estimator of β remains unbiased in small samples under *CL1*, *CL2*, *CL3c*, *CL4*, and *CL5* and normally distributed additionally assuming *CL3a*. It is consistent and approximately normally distributed

under *CL1*, *CL2*, *CL3d*, *CL4*, and *CL5* 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 variance $V(\widehat{\beta}_K|x_{111},...,x_{GMK})$ of $\widehat{\beta}_K$ differs from the usual OLS one under *CL3c* and *CL4*. This cannot be easily expressed without matrix notation even for the binary regression model. Consequently, the usual estimator $\widehat{V}(\widehat{\beta}_k|x_{111},...,x_{GMK})$ 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 $G \to \infty$ and fixed cluster size $M = N/G$, the OLS estimator is asymptotically normally distributed under *CL1*, *CL2*, *CL3d*, *CL4*, and *CL5*

$$\sqrt{G}(\widehat{\beta}_k - \beta_k) \xrightarrow{d} N\left(0, \varsigma^2\right)$$

where ς^2 is not easily expressed without matrix notation. The OLS estimator is therefore approximately normally distributed in samples with a large number of clusters

$$\widehat{\beta}_k \overset{A}{\sim} N\left(\beta_k, Avar(\widehat{\beta}_k)\right)$$

where $Avar(\widehat{\beta}_k) = \varsigma^2/N$ can be consistently estimated with some additional assumptions on higher order moments of $x_{g11},...,x_{gMK}$. For the binary regression, the robust variance estimator is calculated as

$$\widehat{Avar}(\widehat{\beta}_1) = \frac{\sum_{g=1}^{G}\sum_{i=1}^{M}\sum_{j=1}^{M}\widehat{u}_{gi}\widehat{u}_{gj}(x_{gi}-\bar{x})(x_{gj}-\bar{x})}{\left[\sum_{g=1}^{G}\sum_{i=1}^{M}(x_{gi}-\bar{x})^2\right]^2}$$

This so-called *cluster-robust* covariance matrix estimator is a generalization of Huber(1967) and White(1980).[1] 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.

[1] Note: the cluster-robust estimator is not clearly attributed to a specific author.

Note: the cluster-robust covariance matrix is consistent when the number of clusters $G \to \infty$. In practice we should have at least 50 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 gi) and taking all M observations for each drawn cluster. This so-called *block bootstrap* preserves all within cluster correlation. With 20 to 50 clusters, a wild block residual bootstrap-t should be used (Cameron and Miller, 2015).

6 Efficient Estimation with GLS

In some cases, for example with cluster specific random effects, we can estimate β 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.

7 Special Case: Estimating Correct Standard Errors with Random Cluster-specific Effects

Moulton (1986, 1990) studies the bias of the usual OLS standard errors for the special case with random cluster-specific effects. Assume cluster-specific random effects in a bivariate regression:

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

where $u_{gi} = c_g + v_{gi}$ with $\sigma_u^2 = \sigma_c^2 + \sigma_v^2$, $\rho_u = \sigma_c^2/(\sigma_c^2 + \sigma_v^2)$. Then the (cluster-robust) asymptotic variance can be estimated as

$$\widehat{Avar}_{cluster}[\widehat{\beta}_1] = \frac{\widehat{\sigma}_u^2}{\sum_{g=1}^{G}\sum_{i=1}^{M}(x_{gi} - \bar{x})^2}[1 + (M-1)\widehat{\rho}_x\widehat{\rho}_u]$$

where $\widehat{\sigma}_u^2$ is the usual OLS estimator, ρ_x is the within cluster correlation of x. $\widehat{\sigma}^2$, $\widehat{\rho}_u$ and $\widehat{\rho}_x$ are consistent estimators of σ^2, ρ_u and ρ_x, respectively. The robust standard error for the slope coefficient is accordingly

$$\widehat{se}_{cluster}(\widehat{\beta}_1) = \widehat{se}_{ols}(\widehat{\beta}_1)\sqrt{1 + (M-1)\widehat{\rho}_x\widehat{\rho}_u}$$

where $\widehat{se}_{ols}[\widehat{\beta}_1]$ is the usual OLS standard error.

$\sqrt{1 + (M-1)\rho_x\rho_u} > 1$ is called the *Moulton factor* and measures how much the usual OLS standard errors understate the correct standard errors. For example, with cluster size $M = 500$ and intracluster correlations $\rho_u = 0.1$ and $\rho_x = 0.1$, the correct standard errors are 2.45 times the usual OLS ones.

Lessons from the Moulton factor

1. If either the within cluster correlation of the combined error term u is zero ($\rho_u = 0$) or the within cluster correlation of x is zero ($\rho_x = 0$), then the Moulton factor is 1 and the usual OLS standard errors are correct. Both situations generalize to K explanatory variables.

2. If the variable of interest is an *aggregate variable* on the level of the cluster (hence $\rho_x = 1$), the Moulton factor is maximal. This case generalizes to K aggregate explanatory variables:

   $$\widehat{se}_{cluster}(\widehat{\beta}_k) = \widehat{se}_{ols}(\widehat{\beta}_k)\sqrt{1 + (M-1)\widehat{\rho}}$$

   In this situation, we need to correct the standard errors. Alternatively, we could aggregate (average) all variables and run the regression on the collapsed data.

3. If only control variables are aggregated, we better include cluster fixed effects (i.e. dummy variables for the groups) which will take care of the cluster-specific effect. See also the handout on "Panel Data: Fixed and Random Effects".

4. If the variable of interest is not aggregated but has an important cluster specific component (large ρ_x), then including cluster fixed effects may destroy valuable information and we better don't include cluster fixed effects. However, we need to correct the standard errors.

5. If only control variables have an important cluster-specific component, it is better to include cluster fixed effects.

6. If the variable of interest 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.

Standard errors are in practice most easily corrected using the Eicker-Huber-White cluster-robust covariance from section 5 and not via the Moulton factor. Note that we should have at least $G = 50$ clusters 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 17

Load example data

```
webuse auto7.dta
```

Stata reports the cluster-robust covariance estimator clustered for `manufacturer` with the `vce(cluster)` option, e.g.[2]

```
regress price weight, vce(cluster manufacturer)
matrix list e(V)
```

Note: Stata multiplies $\widehat{V}$ with $(N-1)/(N-K-1)\cdot G/(G-1)$ to "correct" for degrees of freedom in small samples. This practice is not based on asymptotic theory but often produces better small sample properties. Stata reports p-values for the t- and F-statistics with $G-1$ degrees of freedom.

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

```
regress price weight, vce(bootstrap, reps(999) cluster(manufacturer))
bootstrap, reps(999) 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)
```

The wild block residual bootstrap-t for the slope coefficient of the variable `weight` is reported by David Roodman's command `boottest`

```
ssc install boottest
regress price weight, vce(cluster manufacturer)
boottest weight=0, reps(99999)
```

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

9 Implementation in R 4.3.1

Load example data

```
library(haven)
auto <- read_dta("http://www.stata-press.com/data/r17/auto7.dta")
```

First, we estimate the regression with the usual command

```
ols <- lm(price~weight, data=auto)
summary(ols)
```

The cluster-robust covariance estimator clustered for `manufacturer` is calculated and reported with the packages `sandwich` and `lmtest`[3]

```
library(sandwich)
library(lmtest)
coeftest(ols, vcov = vcovCL, cluster = ~manufacturer)
```

The following commands are equivalent

```
coeftest(ols, vcov = vcovCL, cluster = ~manufacturer, cadjust=TRUE)
coeftest(ols, vcov = vcovCL(ols, cluster = ~manufacturer))
coeftest(ols, vcov = vcovCL(ols, type="HC1", cluster = ~manufacturer))
```

Note: The above commands multiply $\widehat{V}$ with $(N-1)/(N-K-1)\cdot G/(G-1)$ to "correct" for degrees of freedom in small samples. R reports p-values for the t- and F-statistics with $N-K-1$ degrees of freedom.

We can also estimate a cluster robust covariance using a nonparametric block bootstrap

```
coeftest(ols, vcov = vcovBS, cluster = ~manufacturer, R=999)
```

The wild block residual bootstrap-t for the slope coefficient of the variabel `weight` is calculated by David Roodman's algorithm in `boottest`

```
library(fwildclusterboot)
wild <- boottest(ols, param="weight", clustid=c("manufacturer"),
                 B=99999, type="rademacher", impose_null=TRUE,
                 p_val_type="two-tailed")
summary(wild)
```

[3] 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

Cameron, A. Colin and Douglas L. Miller (2015), A Practitioner's Guide to Cluster-Robust Inference, *Journal of Human Resources*, forthcoming.

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.