Unversitat Pompeu Fabra
Short Guides in Microeconometrics

Kurt Schmidheiny
October 2008

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 (A5a) 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 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

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

where observations belong to a cluster $g = 1, ..., G$ and observations are indexed by $i = 1, ..., N_g$ 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, β is a $(K+1)$-dimensional column vector of parameters, and ε_{ig} is the error term.

Stacking observations within a cluster, we can write

$$y_g = X_g\beta + \varepsilon_g$$

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

$$y = X\beta + \varepsilon$$

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

The data generation process (dgp) is fully described by the following set of assumptions:

A1: Linearity

$y_i = x'_{ig}\beta + \varepsilon_{ig}$ and $E(\varepsilon_{ig}) = 0$

A2: Independence

c) $(X_g, y_g)_{g=1}^G$ independently distributed

A2c means that the observations in one cluster are independent from the observations in all other clusters.

A3: Exogeneity

a) $\varepsilon_{ig}|X_g \sim N(0, \sigma^2_{ig})$

b) $\varepsilon_{ig} \perp X_g$ and $E(\varepsilon_{ig}) = 0$ (independent)

c) $E(\varepsilon_{ig}|X_g) = 0$ (mean independent)

d) $Cov(X_g, \varepsilon_{ig}) = 0$ and $E(\varepsilon_{ig}) = 0$ (uncorrelated)

Note that the error term ε_{ig} is assumed unrelated to the explanatory variables (X_g) of all observations within its cluster.

A4: Identifiability

$$rank(X) = K + 1 < N$$

A5: Error Variance

c) $V(\varepsilon_{ig}|X_g) = \sigma^2_{ig} < \infty$, for all i, g
$Cov(\varepsilon_{ig}, \varepsilon_{jg}|X_g) = \rho_{ijg}\sigma_{ig}\sigma_{jg} < \infty$, for all $i \neq j, g$

A5c means that the error terms are correlated within clusters (clustered) and have different variances (heteroscedastic).

A6: Variance of explanatory variables

a) $V(X) = E(X'X)$ is positive definite and finite

b) $plim(\frac{1}{N}X'X) = Q_{XX}$ is positive definite and finite

The variance-covariance of the vector of error terms in the whole sample is under A2 and A5

$$\Omega = V(\varepsilon|X) = E(\varepsilon\varepsilon'|X)$$
$$= \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{pmatrix}$$

where, for example,

$$\Omega_1 = V(\varepsilon_1|X_1) = E(\varepsilon_1\varepsilon_1'|X_1)$$
$$= \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1N_1}\sigma_1\sigma_{N_1} \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2N_1}\sigma_2\sigma_{N_1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N_1}\sigma_1\sigma_{N_1} & \rho_{2N_1}\sigma_2\sigma_{N_1} & \cdots & \sigma_{N_1}^2 \end{pmatrix}$$

is the variance covariance of the error terms within cluster $g = 1$.

3 A Special Case: Cluster Specific Random Effects

Suppose as Moulton(1986) that the error term ε_{ig} consists of a cluster specific random effect α_g and an individual effect ν_{ig}

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

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

$$V(\nu_{ig}|X_g) = \sigma^2_\nu$$
$$Cov(\nu_{ig}, \nu_{jg}|X_g) = 0,\ i \neq j$$

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

$$V(\alpha_g|X_g) = \sigma^2_\alpha$$
$$Cov(\alpha_g, \nu_{ig}|X_g) = 0$$

The cluster specific effect α_g is under A3 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

$$\Omega_g = V(\varepsilon_g|X_g) = \begin{pmatrix} \sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{pmatrix}$$

where $\sigma^2 = \sigma^2_\alpha + \sigma^2_\nu$ and $\rho = \sigma^2_\alpha/(\sigma^2_\alpha + \sigma^2_\nu)$. In a less restrictive version, σ^2_g and ρ_g are allowed to be cluster specific.

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 β can be estimated with OLS as

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

The OLS estimator of β remains unbiased (under A1, A2c, A3c, A4, A5c and A6) and normally distributed (additionally assuming A3a) in small samples. It is consistent and approximately normally distributed (under A1, A2c, A3d, A4, A5c and A6b) 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 the Covariance of the OLS Estimator

The small sample covariance matrix of $\hat{\beta}_{OLS}$ is under A3c and A5c

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

and differs from usual OLS where $V = \sigma^2(X'X)^{-1}$. Consequently, the usual estimator $\hat{V} = \hat{\sigma}^2(X'X)^{-1}$ 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 $G \rightarrow \infty$, the OLS estimator is asymptotically normally distributed under A1, A2, A3d, A4, A5c and A6b

$$\sqrt{G}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, Q^{-1}\Sigma Q^{-1}\right)$$

The OLS estimator is therefore approximately normally distributed in samples with a large number of clusters

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

where $V = G^{-1}Q^{-1}\Sigma Q^{-1}$ can be consistently estimated as

$$\hat{V} = (X'X)^{-1} \left(\sum_{g=1}^{G} X_g' e_g e_g' X_g\right) (X'X)^{-1}$$

with $e_g = y_g - X_g\hat{\beta}_{OLS}$.

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.

Note: the cluster-robust covariance matrix is consistent when the number of clusters $G \rightarrow \infty$ and the number of observations per cluster N_g is fixed. In practice this requires a sample with many clusters (50 or more) and relatively small number of observations per cluster.

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 Estimation with Cluster Specific Random Effects

In the cluster specific random effects model, the error covariance matrix Ω only depends on the two parameters ρ and σ. These two parameters can be consistently estimated in samples with many clusters. We could plug these estimates into Ω to estimate the correct covariance $\hat{V}$ for the OLS estimator $\hat{\beta}_{OLS}$.

[1] 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`

However, if we are willing to assume cluster specific random effects, we can directly 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 Implementation in Stata 10.0

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

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

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

We can also estimate a heteroscedasticity robust covariance using a nonparametric block bootstrap. For example,

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

or

```
bootstrap, rep(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)
```

[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.

References

Cameron, A. C. and P. K. Trivedi (2005), Microeconometrics: Methods and Applications, Cambridge University Press. Sections 24.5.

Wooldridge, J. M. (2002), Econometric Analysis of Cross Section and Panel Data. MIT Press. Sections 7.8 and 11.5.

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, 221223.

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.