

ARE213

Econometrics

Spring 2006 UC Berkeley Department of Agricultural and Resource Economics

## ORDINARY LEAST SQUARES IV:

## CLUSTERING AND VARIANCE ESTIMATION (W 6.3.4, Mo)

When we looked at the standard linear model

$$Y_i = X_i' \beta + \varepsilon_i,$$

where  $X_i$  is an  $L$ -dimensional column vector, or in matrix notation,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

we assumed we had independent observations. Often that is not quite true. In general this makes progress difficult, but progress can be made if we impose some additional structure. Suppose that the pairs  $(Y_i, X_i)$  are *clustered*. Let  $S_i$  be index for the cluster, so that with  $K$  clusters  $S_i \in \{1, \dots, K\}$ . Within each cluster the  $(Y_i, X_i)$  are correlated, but  $(Y_i, X_i)$ 's from different clusters are independent. Clusters could be states, or classrooms, or any other grouping that could be expected to lead to a particular form of dependencies. To do asymptotics we assume that the number of observations per cluster is fixed and the number of clusters increases. Let us initially also assume that the number of observations per cluster is the same for all clusters, and equal to  $M$ . More generally the sample size in cluster or group  $k$  is  $M_k$ . The total sample size is  $N = \sum_{k=1}^K M_k$ , equal  $M \cdot K$  in the special case with a constant group size. This structure will be seen to greatly affect standard errors if the variable of interest varies only between clusters.

It is useful to introduce some additional notation and give some preliminary results. Let  $\mathbf{Z}$  be the  $N \times K$  matrix of group or cluster indicators with typical element  $Z_{ij} = 1\{S_i = j\}$ . For example, with three clusters, ten observations, of which the first two are from cluster 1,

the next five are from cluster two, and the last three are from cluster three, we would have

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

If  $\iota_N$  is the  $N$ -dimensional vector with all elements equal to one, then  $\mathbf{Z}'\iota_N$  gives a  $K$  vector with the  $k$ th element equal to  $M_k$ , the group size of cluster  $k$ . With  $\mathbf{Y}$  an  $N$ -dimensional vector,  $(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y})$  is the  $K$  vector with group means:

$$((\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y}))_k = \sum_{i=1}^N 1\{S_i = k\} \cdot Y_i / M_k.$$

Further more, keeping in mind that in general for a matrix  $\mathbf{X}$ ,  $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$  is the projection of  $\mathbf{Y}$  on  $\mathbf{X}$ , we have that  $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y})$  is the  $N$  vector with each element equal to the group mean. Finally,  $\mathbf{Y} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{Y}) = (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{Y}$  is the vector of deviations of  $\mathbf{Y}$  from group means.

Clustering can be an issue in much more general problems, but here we look at a special case, the linear model with clustering. Suppose

$$Y_i = X_i'\beta + \varepsilon_i.$$

Let  $\varepsilon$  be the  $N$  vector with  $i$ th element equal to  $\varepsilon_i$ . Suppose that

$$E[\varepsilon] = 0,$$

$$E[\varepsilon\varepsilon'] = \sigma^2 \cdot ((1 - \rho) \cdot I_N + \rho \cdot \mathbf{Z}\mathbf{Z}').$$

For the previous example we have

$$\mathbf{ZZ}' = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

Since the diagonal elements of  $\mathbf{ZZ}'$  are equal to one, this implies that the variance of  $\varepsilon_i$  is equal to  $\sigma^2$ .

An alternative way to think about this error components structure is to think of  $\varepsilon_i = \eta_i + \nu_i$ , where  $\eta_i$  and  $\nu_i$  are independent,  $\eta_i$  has variance  $(1 - \rho) \cdot \sigma^2$ , and  $\nu_i$  varies only between clusters, not within, and has variance  $\rho \cdot \sigma^2$ .

OLS is still consistent here, although it is not efficient. An efficient estimator could be based on Generalized Least Squares (GLS) although we do not pursue this here. See for more discussion on GLS Wooldridge. The standard ols variance for the least squares estimator for  $\beta$ ,

$$\hat{\beta}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}),$$

is

$$\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}. \tag{1}$$

Under the variance structure implied by the model the true variance is

$$\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot (I_L + \rho \cdot (\mathbf{X}'\mathbf{ZZ}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - I_L)), \tag{2}$$

where  $L$  is the number of regressors in  $X$ . Kloek (1981) considers a simplification where all regressors  $X$  are fixed within the groups, and the group sizes are all equal to  $M$ . In that case the true variance simplifies to

$$\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot (1 + (M - 1) \cdot \rho). \quad (3)$$

Moulton suggests that in cases where there are different group sizes using this correction with the average group size  $\bar{M} = \sum_{k=1}^K M_k / K$  may still give a pretty good approximation. More specifically, even if some of the regressors vary within groups it may still give a good approximation to the standard errors of the regressors that are fixed within groups.

To estimate  $\sigma^2$  and  $\rho$ , first calculate the residuals from the ols regression ignoring any clustering:

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

(This is not necessarily efficient, but  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$  is consistent for  $\beta$  since the clustering only affects the variance.) Then estimate the variance parameter  $\sigma^2$  as

$$\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} / (N - L).$$

The degree of freedoms subtracted here are the number of regressors in  $X$ . Since the asymptotics is for  $N$  going to infinity, and  $L$  staying fixed, this does not matter. You could also just divide by  $N$ .

Next, we need to estimate  $\rho$ . To estimate  $\rho$  we first subtract from each residual the mean residual within the group using the projection matrix:

$$\tilde{\varepsilon} = (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \hat{\varepsilon}.$$

These residuals  $\tilde{\varepsilon}$  are approximately uncorrelated within clusters as well as between clusters. Next we estimate the variance of this residual:

$$\tilde{\sigma}^2 = \tilde{\varepsilon}'\tilde{\varepsilon} / (N - K).$$

Now it is important to subtract the degrees of freedom,  $K$ , for the  $K$  means that were subtracted from the residuals. Because  $K$  goes to infinity with the sample size, this will make an important difference, unlike in earlier discussions where the degrees of freedom adjustment was based on a small, fixed number of covariates.

Let us consider the expectation of  $\tilde{\varepsilon}\tilde{\varepsilon}'$ , ignoring the difference between  $\hat{\varepsilon}$  and  $\varepsilon$  (which will be small in large samples):

$$\begin{aligned} \mathbb{E}[\tilde{\varepsilon}\tilde{\varepsilon}'] &= \mathbb{E} \left[ (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \hat{\varepsilon}\hat{\varepsilon}' (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \right] \\ &\approx \mathbb{E} \left[ (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \varepsilon\varepsilon' (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \right] \\ &= (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \sigma^2 \cdot ((1 - \rho) \cdot I_N + \rho \cdot \mathbf{Z}\mathbf{Z}') (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') \\ &= \sigma^2 \cdot (1 - \rho) \cdot (I_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'). \end{aligned}$$

The trace of this matrix, that is the sum of the diagonal elements of this matrix is equal to  $\sigma^2 \cdot (1 - \rho) \cdot (N - K)$ . Hence,  $\tilde{\sigma}^2$  estimates  $\sigma^2 \cdot (1 - \rho)$ , so that we can estimate  $\rho$  as

$$\hat{\rho} = \frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\hat{\sigma}^2},$$

which can be substituted into (3).

Another approach that works well in the Klock case with constant group sizes and only aggregate regressors is the following. First estimate the group means of the outcome:

$$\bar{Y}_k = \frac{1}{M} \sum_{i=1}^M Y_{ik}.$$

Then regress

$$\bar{Y}_k = X_k'\beta + \bar{\varepsilon}_k,$$

using the  $K$  observations, one per cluster. This will give the same estimates as ols on the big  $N$  observation data set, as well as correct standard errors. More generally, divide the regressors into two parts,  $(X_i = (W_i, V_i))$ , where  $W_i$  are the regressors that vary within the clusters and  $V_i$  are those that vary only between clusters, and let  $Z_i$  denote the  $K$  vector of cluster dummies,  $Z_{ik} = 1\{S_i = k\}$ . Moulton suggests for running the regression

$$Y_i = W_i' \beta_W + \delta' Z_i + \eta_i.$$

Then in the second stage run the regression

$$\hat{\delta}_k = V_k' \beta_X + \varepsilon_k,$$

with the cluster specific variables.

Let us see how this works out in practice. I took the census data from Angrist and Krueger to estimate a regression with both individual level education and the average of state education levels. The idea is to see if education of those around a person affect their earnings as well through interactions. The regression function is:

$$\log(\text{earnings})_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{state} - \text{educ}_i + \varepsilon_i.$$

I use about 329,509 observations, 51 clusters, and so on average 6,461 observations per cluster. The estimated correlation coefficient is very small, approximately 0.0005.

I estimate the regression using least squares and calculate the standard errors in three ways. First the conventional ols standard errors. Second the correct standard errors given in (2), and finally the standard errors suggested by Klock (3).

Note that the Klock standard errors (based on (3)) are pretty good for the average education variable, but not for the individual level education variable. The Moulton (correct) standard errors (based on (2)) are very close to ols for the individual regressor, but very different from ols for the aggregate regressor.

Table 1: ESTIMATES AND STANDARD ERRORS

	intercept	own education	average state education
estimate	4.2584	0.0656	0.0665
ols	(0.0542)	(0.0011)	(0.0045)
Moulton	(0.1345)	(0.0011)	(0.0110)
Kloek	(0.1105)	(0.0022)	(0.0091)

Next, I estimate the coefficient on average education using the Moulton-Kloek suggestion of first estimating state dummies. This leads to

Table 2: SECOND STAGE KLOEK-MOULTON REGRESSION

	intercept	average state education
estimate	4.3510	0.0585
s.e.	(0.1606)	(0.0129)

Although this does not give results identical to those of the single regression approach, the estimates and standard errors are fairly close, and certainly the standard errors are no longer misleading the way they were if you use conventional ols standard errors.

#### REFERENCES

KLOEK, T., (1981), "OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," *Econometrica*, Vol. 49, No. 1, 205-207.

MOULTON, B., (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 334-338.

MOULTON, B., AND W. RANDOLPH, (1989) "Alternative Tests of the Error Component Model," *Econometrica*, Vol. 57, No. 3, 685-693.