Lecture 9: Standard errors, clustering and bootstrapping

Søren Leth-Petersen and Daniel le Maire
Department of Economics
University of Copenhagen

October 6, 2022

Contents

1 Inference for the linear model

Most focus is typically on obtaining unbiased or consistent point estimates, i.e. estimating β. However, to interpret this point estimate we need that our estimated standard errors are unbiased to avoid misleading inference.

In these notes, much of our focus will be on the case where observations are grouped into clusters, where the error terms are correlated within clusters, but uncorrelated across clusters. When making standard errors robust to heteroscedasticity, it usually increases the standard errors, but not a lot. When controlling for intra-cluster correlation, it is not unusual that standard errors becomes several times larger.

We have already considered cluster-robust standard errors for the case of individual fixed effects. In this case, the individual was treated as a cluster to take account of (within-individual) serial correlated errors. However, clustering is much more general problem, which arise in both panel data and cross-sectional settings. Throughout these notes we will focus on clustering in a cross-sectional data set. Furthermore, the focus in the notes is on how to correct the standard errors of parameter estimates and consequently we will not look into how to obtain more efficient parameter estimates using feasible generalized least squares (FGLS).

We will begin these notes by deriving the standard errors under homoscedasticity and heteroscedasticity. Next, we will consider two ways of correcting the standard errors in case of clustering. Finally, we look at how to use bootstrapping to estimate standard errors.[1]

[1] These notes mainly build on Angrist and Pischke (2009, chapter 8), Cameron and Miller (2015). The final part on bootstrapping also draws on Horowitz (2001) and Poi (2004).

2 Classical and heteroscedasticity robust standard errors

Let $y_i = \mathbf{x}_i\beta + u_i$, where y_i is the dependent variable for individual $i = 1, ..., N$, $\mathbf{x}_i$ is a $1 \times K$ vector of explanatory variables for individual i (including a constant), β is the $K \times 1$ vector of parameters, and u_i is the error term. Under the assumption of $E(u_i|\mathbf{x}_i) = 0$ and the usual rank condition, the OLS estimator is given by

$$\begin{aligned}\hat{\beta} &= \left[\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\right]^{-1} \sum_{i=1}^{N} \mathbf{x}_i' y_i \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}\end{aligned}$$

where $\mathbf{X}$ is the stacked $\mathbf{x}_i$ with dimension $N \times K$.

To derive the variance-covariance matrix, we insert $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\mathbf{X}\beta + \mathbf{u}) \Leftrightarrow \\ \hat{\beta} - \beta &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\end{aligned}$$

Using this, we can write the variance of $\hat{\beta}$ as

$$\begin{aligned}Avar\left(\hat{\beta}\right) &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\left((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\right)'\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right]\end{aligned}$$

Conditional on $\mathbf{X}$, we have

$$Avar\left(\hat{\beta}\right) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' E(\mathbf{u}\mathbf{u}'|\mathbf{X}) \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \tag{1}$$

Under homoscedasticity, we assume that[2]

$$E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2 \mathbf{I}_T \tag{2}$$

Using this, equation (**??**) gives

$$\begin{aligned}Avar\left(\hat{\beta}\right) &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\sigma^2\mathbf{I}_T\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\end{aligned} \tag{3}$$

where we estimate σ^2 by $\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-K}$ where $\hat{u}_i = y_i - \mathbf{x}_i\hat{\beta}$.[3] The standard errors of $\hat{\beta}$ can be found as the square root of the diagonal of the r.h.s. of

[2] This is equivalent to assuming $E\left(u_i^2\mathbf{x}_i'\mathbf{x}_i\right) = \sigma^2 E\left(\mathbf{x}_i'\mathbf{x}_i\right)$ as in OLS.3.

[3] Since we deal with the asymptotic variance, $\hat{Avar}\left(\hat{\beta}\right)$, the degrees of freedom is not really necessary.

equation (**??**). In these notes, we will refer to these standard errors as *classical standard errors*.

Heteroscedasticity implies that the variance of the error term is non-constant, but independent. In this case,

$$E\left(\mathbf{uu}'|\mathbf{X}\right)=\mathbf{\Omega}=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & 0 & \sigma_N^2\end{bmatrix}$$

Using this assumption, White (1980a) showed that the following asymptotic variance is robust to heteroscedasticity of unknown form

$$\begin{aligned}Avar\left(\hat{\beta}\right) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{N}\sigma_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad (4)\end{aligned}$$

where we estimate σ_i^2 by $\hat{u}_i^2$.[4] We will refer to the standard errors which can be computed based on equation (**??**) as *heteroscedasticity robust standard errors*.

We should never rely solely on the classical standard errors. In fact, if you view a regression as a linear approximation to the conditional expectation function $E\left(y_i|\mathbf{x}_i\right)$ then, as White (1980b) noticed, a true non-linear conditional expectation function would imply heteroscedastic residuals even if you are prepared to assume that the conditional variance $Var\left(y_i|\mathbf{x}_i\right)$ is constant. To see this, write

$$\begin{aligned}E\left[\left(y_i-\mathbf{x}_i\beta\right)^2|\mathbf{x}_i\right] &= E\left[\left(y_i-E\left(y_i|\mathbf{x}_i\right)+E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)^2|\mathbf{x}_i\right]\\ &= E\left[\begin{array}{c}\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)^2+\left(E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)^2\\ +2\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)\left(E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)\end{array}\middle|\mathbf{x}_i\right]\\ &= E\left[\begin{array}{c}\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)^2+\left(E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)^2\\ +2\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)E\left(y_i|\mathbf{x}_i\right)-2\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)\mathbf{x}_i\beta\end{array}\middle|\mathbf{x}_i\right]\\ &= \left[\begin{array}{c}E\left[\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)^2|\mathbf{x}_i\right]+E\left[\left(E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)^2|\mathbf{x}_i\right]\\ +2\underbrace{E\left[\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)E\left(y_i|\mathbf{x}_i\right)|\mathbf{x}_i\right]}_{=0}-2\underbrace{E\left[\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)\mathbf{x}_i\beta|\mathbf{x}_i\right]}_{=0}\end{array}\right]\\ &= Var\left(y_i|\mathbf{x}_i\right)+\left(E\left(y_i|\mathbf{x}_i\right)-\mathbf{x}_i\beta\right)^2\end{aligned}$$

where we have used that $\left(y_i-E\left(y_i|\mathbf{x}_i\right)\right)$ is mean independent of $\mathbf{x}_i$. Therefore, even if you are willing to assume that $Var\left(y_i|\mathbf{x}_i\right)$ is constant, the residual variance will be varying if the true model is not linear.

- Typically, the heteroscedasticity robust standard errors will be larger than the classical standard errors.

[4] It is actually the step where we replace σ_i^2 by $\hat{u}_i^2$, which is the key result in White (1980a).

- Asymptotically, classical and heteroscedasticity robust standard errors are correct under the appropriate assumptions, but both suffer from finite sample bias, that will tend to make them too small in small samples.

- It is primarily the heteroscedasticity robust standard errors that can have a large bias. Chesher and Jewitt (1987) show that if there is not "too much" heteroscedasticity, the heteroscedasticity robust standard errors will even in "fairly large" samples be downward biased.

- Part of the problem is that no correct degrees of freedom correction exists for the heteroscedasticity robust standard errors.[5] Recall that the degrees of freedom correction used under homoscedasticity correct for the OLS residuals systematically being too close to zero. The reason for this is that we have K restrictions on the OLS residuals, $\sum_{i=1}^{N} \mathbf{x}_{ij}\hat{u}_i = 0$ for the $j = 1, .., K-1$ explanatory variables and $\sum_{i=1}^{N} \hat{u}_i = 0$, whereby only $N - K$ residuals are free to vary.

Since no correct degrees of freedom correction exists for the heteroscedasticity robust standard errors, it can easily be the case that the classical OLS standard errors are largest. Below, we will discuss when cluster robust standard errors are called for. When this is not required, we should select the largest of the classical and heteroscedasticity robust standard errors.

3 Clustering

When errors are positively correlated within a cluster (or a non-overlapping group) then an additional observation in the cluster no longer provides a completely independent piece of new information. Therefore, we cannot just use the classical or heteroscedastic standard errors as they do not take the intra-cluster correlation into account.

The standard example of clustering is a case where the left hand side variable is an individual outcome and where at least one of the explanatory variables is an aggregate variable. For example, if the outcome is the individual probability of finding a job, one relevant aggregate variable could be the local unemployment rate. Observations within a cluster can be thought to be correlated as a result of an unobserved cluster effect. To be more precise, an explanatory variable, which is aggregated, will not necessarily lead to a clustering problem: If the inclusion of this variable mops up all of the cluster-specific effect such that the remaining error-term only has individual variation, we will not need to compute cluster-robust standard errors. Clearly, this is highly unlikely to be the case as this would imply that we have the right model in terms of functional form for the aggregate level and that our aggregate explanatory variable is not measured with error.

[5] MacKinnon and White (1985) discuss different type of corrections for the residual variance estimator, for example $\sigma_i^2 = \frac{N}{N-K}\hat{u}_i^2$, instead of $\sigma_i^2 = \hat{u}_i^2$, which White (1980a) used.

For example, imagine that you conduct an experiment where you want to measure the effect of a new teaching program. Within 20 different schools (each with 4 classes with 25 students in each class) you randomize which classes should have the program and not, so that 50 percent of all classes within each school are given the program. You then collect test scores for all students in all the classes in the 20 schools. Now the experimental variation is at the class level and not at the student level. This means that even if we have data on 2000 students it is only the class level that is useful for estimating the effect of the program. Actually it is as if we only have 80 observations corresponding to the number of classes. Using the individual level data set with 2000 observations we should therefore cluster at the class level, i.e. allow covariances between students within the same classes to be unrestricted.

3.1 The Moulton factor

We want to study the relationship between the classical standard errors and clustered standard errors for the simplest case where we only have $\mathbf{x}$-variables which are constant within each cluster. The equation of interest for member m of the cluster is given by

$$y_{gm} = \beta_0 + \mathbf{x}_{gm}\beta_1 + v_{gm} \tag{5}$$

When stacking equation (**??**), we will make use of the fact that all members of the cluster g share the exact same values of the explanatory variables as member m, that is $x_{gj} = x_{gm}$ for all $j = 1, 2, ..., M_g$ where M_g is the number of members (or observations) in cluster g. Hence, when we stack $\mathbf{x}_{gm}$, we will use the $M_g \times K$ matrix $\mathbf{x}_g = \mathbf{j}_{M_g}\mathbf{x}_{gm}$ where $\mathbf{j}_{M_g}$ is a $M_g \times 1$ vector of ones. Stacking observations within a cluster we can write

$$y_g = \beta_0 + \mathbf{j}_{M_g}\mathbf{x}_{gm}\beta_1 + v_g$$

We assume that the residual has a random-effects group structure

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

where c_g is a random unobservable group component and u_{gm} is a mean-zero individual-level component. The implication of the random-effects structure is that errors are equicorrelated within clusters. This is a suitable assumption when the ordering of the observations within clusters is irrelevant. When the cluster is a geographical unit this may be a reasonable assumption.

Given this error-structure, the variance of the composite error-term is

$$\sigma_v^2 = \sigma_c^2 + \sigma_u^2$$

The variance-covariance matrix is

$$E\left(\mathbf{uu}'|\mathbf{X}\right) = \mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{\Omega}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{\Omega}_G \end{bmatrix}$$

where

$$\begin{aligned}\boldsymbol{\Omega}_g &= \begin{bmatrix} \sigma_c^2+\sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2+\sigma_u^2 & \cdots & \sigma_c^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_c^2 & \sigma_c^2 & \cdots & \sigma_c^2+\sigma_u^2 \end{bmatrix} \\ &= \left[\mathbf{I}_{M_g}\sigma_u^2 + \mathbf{j}_{M_g}\mathbf{j}'_{M_g}\sigma_c^2\right] \end{aligned} \tag{6}$$

where $\mathbf{I}_{M_g}$ is a $M_g \times M_g$ identity matrix and $\mathbf{j}_{M_g}$ is a $M_g \times 1$ vector of ones.

We will continue by computing the variances under the assumption of non-stochastic x's. Replacing the $\boldsymbol{\Omega} = E(\mathbf{u}\mathbf{u}'|\mathbf{X})$, the variance is given by

$$Avar\left(\hat{\beta}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{7}$$

First consider the matrix $\mathbf{X}'\mathbf{X}$

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \sum_{g=1}^{G}\left(\mathbf{j}_{M_g}\mathbf{x}_{gm}\right)'\mathbf{j}_{M_g}\mathbf{x}_{gm} \\ &= \sum_{g=1}^{G}\mathbf{x}'_{gm}\mathbf{j}'_{M_g}\mathbf{j}_{M_g}\mathbf{x}_{gm} \\ &= \sum_{g=1}^{G}M_g\mathbf{x}'_{gm}\mathbf{x}_{gm} \end{aligned} \tag{8}$$

Next, consider the matrix $\mathbf{X}'\boldsymbol{\Omega}\mathbf{X}$

$$\begin{aligned}\mathbf{X}'\boldsymbol{\Omega}\mathbf{X} &= \sum_{g=1}^{G}\mathbf{x}'_{gm}\mathbf{j}'_{M_g}\boldsymbol{\Omega}_g\mathbf{j}_{M_g}\mathbf{x}_{gm} \\ &= \sum_{g=1}^{G}\mathbf{x}'_{gm}\mathbf{j}'_{M_g}\left[\mathbf{I}_{M_g}\sigma_u^2 + \mathbf{j}_{M_g}\mathbf{j}'_{M_g}\sigma_c^2\right]\mathbf{j}_{M_g}\mathbf{x}_{gm} \\ &= \sum_{g=1}^{G}\left(\mathbf{x}'_{gm}\mathbf{j}'_{M_g}\mathbf{j}_{M_g}\sigma_u^2\mathbf{x}_{gm} + \mathbf{x}'_{gm}\mathbf{j}'_{M_g}\mathbf{j}_{M_g}\mathbf{j}'_{M_g}\sigma_c^2\mathbf{j}_{M_g}\mathbf{x}_{gm}\right) \\ &= \sum_{g=1}^{G}M_g\left(\sigma_u^2 + M_g\sigma_c^2\right)\mathbf{x}'_{gm}\mathbf{x}_{gm} \\ &= \sigma_v^2\sum_{g=1}^{G}M_g\left(\frac{\sigma_u^2}{\sigma_c^2+\sigma_u^2} + M_g\frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}\right)\mathbf{x}'_{gm}\mathbf{x}_{gm} \\ &= \sigma_v^2\sum_{g=1}^{G}M_g\left(1 + (M_g-1)\frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}\right)\mathbf{x}'_{gm}\mathbf{x}_{gm} \\ &= \sigma_v^2\sum_{g=1}^{G}M_g\left(1 + (M_g-1)\rho_v\right)\mathbf{x}'_{gm}\mathbf{x}_{gm} \end{aligned} \tag{9}$$

where $\rho_v \equiv \frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}$ is called the intra-class correlation coefficient.

Inserting equations (**??**) and (**??**) in the variance formula (**??**), we obtain

$$Avar\left(\hat{\beta}_1\right) = \sigma_v^2 \left(\sum_{g=1}^{G} M_g \mathbf{x}_{gm}' \mathbf{x}_{gm}\right)^{-1} \sum_{g=1}^{G} M_g \left(1 + \left(M_g - 1\right)\rho_v\right) \mathbf{x}_{gm}' \mathbf{x}_{gm} \left(\sum_{g=1}^{G} M_g \mathbf{x}_{gm}' \mathbf{x}_{gm}\right)^{-1}$$

Now, suppose the cluster sizes are equal, i.e. $M_g = M$. Then, we have that

$$Avar\left(\hat{\beta}_1\right) = \left(1 + \left(M - 1\right)\rho_v\right) \sigma_v^2 \left(\sum_{g=1}^{G} M \mathbf{x}_{gm}' \mathbf{x}_{gm}\right)^{-1} \tag{10}$$

Comparing this cluster-robust variance with the classical OLS variance

$$Avar\left(\hat{\beta}_1\right) = \sigma_v^2 \left(\sum_{g=1}^{G} M_g \mathbf{x}_{gm}' \mathbf{x}_{gm}\right)^{-1} \tag{11}$$

we can compute the ratio between the two variances in equations (**??**) and (**??**)[6] as $1 + \left(M - 1\right)\rho_v$. The square-root of this is called the Moulton factor after Moulton (1986) since

$$se_{cluster}\left(\hat{\beta}_1\right) = se_{classical}\left(\hat{\beta}_1\right)\sqrt{1 + \left(M - 1\right)\rho_v}$$

- The Moulton factor tells us how much the precision of the standard errors is overstated when ignoring the intra-cluster correlation.
- For a fixed number of observations, N, a higher M implies fewer clusters and since there is no dependence between clusters and only dependence within each cluster, this increases the overall dependence in the data. In other words, with the classical standard errors, each additional observation adds new and independent information, but in reality additional observations within a group does not add so much information.
- A higher intra-class correlation ρ_v increases the Moulton factor since this implies that additional observations within a group provide less new information and this is not acknowledged using the classical standard errors.

The Moulton factor for the case where the covariate x_{gm} varies at the individual level and the cluster size also varies between clusters is given by

$$se_{cluster}\left(\hat{\beta}_1\right) = se_{classical}\left(\hat{\beta}_1\right)\sqrt{1 + \left(\frac{Var\left(M_g\right)}{\bar{M}} + \bar{M} - 1\right)\rho_x\rho_v} \tag{12}$$

where $\bar{M} \equiv \frac{1}{G}\sum_{g=1}^{G} M_g$ and $\rho_x \equiv \frac{\sum_{g=1}^{G}\sum_{m=1}^{M_g}\sum_{j=1}^{M_g} 1[m\neq j](x_{gm}-\bar{x})(x_{gj}-\bar{x})}{V(x_{gm})\sum_{g=1}^{G} M_g(M_g-1)}$ is the intra-cluster correlation of x_{gm}.

Notice that if x_{gm} is uncorrelated within clusters, there is no clustering problem and the Moulton factor is 1. Hence, clustering is likely to be a larger problem with covariates being fixed within clusters.

[6] Where we for simplicity set $M_g = M$ for all g.

3.2 Cluster-corrected standard errors

Besides using the Moulton factor to correct estimated standard errors, it is also possible to compute cluster-robust standard errors directly. With the cluster structure, errors are only correlated within-cluster and not across clusters. Hence, the errors have a block-diagonal structure. The within-cluster variance-covariance matrix of the residuals is[7]

$$\mathbf{\Omega}_g = E\left(\mathbf{u}_g\mathbf{u}_g'|\mathbf{x}_g\right) = \begin{bmatrix} \sigma_{g11}^2 & \sigma_{g12}^2 & \cdots & \sigma_{g1M}^2 \\ \sigma_{g21}^2 & \sigma_{g22}^2 & \cdots & \sigma_{g2M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{gM1}^2 & \sigma_{gM2}^2 & \cdots & \sigma_{gMM}^2 \end{bmatrix} \tag{13}$$

Notice that within each cluster no restrictions are imposed on the type of correlation. This is in contrast to the variance-covariance matrix of the residuals in equation (**??**), where the restriction of equicorrelated errors was imposed.

Inserting this in equation (**??**), we arrive at

$$Avar\left(\hat{\beta}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{x}_g'E\left(\mathbf{u}_g\mathbf{u}_g'|\mathbf{x}_g\right)\mathbf{x}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{14}$$

To estimate this variance, $E\left(\mathbf{u}_g\mathbf{u}_g'|\mathbf{x}_g\right)$ in equation (**??**) is replaced by $\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g$, where $\hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\hat{\beta}$. At first sight, this variance looks quite similar to the heteroscedasticity robust variance in equation (**??**). However, whereas we are summing over individual observations in the latter case, we are summing over whole clusters in equation (**??**). This difference reflects that whereas observations are independent in the case of heteroscedasticity such that the variance-covariance matrix of the residuals is a diagonal matrix, observations within a cluster are dependent as is also apparent from the variance-covariance matrix in equation (**??**).

This is a consistent cluster robust variance estimator if

$$G^{-1}\sum_{g=1}^{G}\mathbf{x}_g'\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'\mathbf{x}_g - G^{-1}\sum_{g=1}^{G}E\left(\mathbf{x}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{x}_g\right) \xrightarrow{p} \mathbf{0} \text{ as } G \to \infty$$

Why is the cluster robust variance estimator only consistent as $G \to \infty$, and not $N \to \infty$? To begin with notice that residual variance covariance matrix $\mathbf{uu}'$ in equation (**??**) has dimension $N \times N$. Under homoscedasticity there is no correlation between u_i and u_j for $i \neq j$ and the matrix $\mathbf{uu}'$ is a diagonal matrix. Furthermore, since $E\left(u_i^2\right)$ is the same for all i, we can estimate this by taking the average of u_i^2 for $i = 1, ..., N$. Therefore, we are averaging over N observations and in order to use the Law of Large Numbers we let $N \to \infty$. This way, the classical standard errors are consistent as $N \to \infty$. Compare this with equation (**??**) where $\mathbf{u}_g\mathbf{u}_g'$ is $M_g \times M_g$ and where we cannot set any cells

[7] To ease the notation, we have avoided the subscript g of M_g in the last row and column in the matrix. However, we should think of Ω_g being $M_g \times M_g$.

to zero within a cluster. Replacing $E\left(\mathbf{u}_g\mathbf{u}_g'\right)$ with $\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'$ is likely to be a very poor estimate. However, since we have G clusters, we can average this matrix over the G clusters.[8] To use the Law of Large Numbers for cluster standard errors, we need that $G \to \infty$.

The cluster-robust variance estimator in equation (**??**) and the Moulton factor in equation (**??**) are both asymptotic results for $G \to \infty$. The usual rule of thumb is that these formulas work with 50 or more clusters. With a small number of clusters ($G < 50$) or very unbalanced cluster sizes M_g, inference using cluster-robust errors can be incorrect and worse than the classical or heteroscedasticity robust standard errors.

- As with heteroscedasticity-robust standard errors, there exist no correction, which will make cluster robust standard errors unbiased. In practice the finite-sample adjustment factor $\frac{N-1}{N-K}\frac{G}{G-1}$ is often used (as it is implemented in STATA). In SAS, the finite sample correction is $\frac{G}{G-1}$ and this simpler correction term is also used in STATA for non-linear models.
- A second general problem with few clusters also arise:
  - Recall, that with OLS under homoscedasticity, we need to use the t-distribution rather than the standard normal distribution in small samples since we replace the variance of the error term with the estimated counterpart when estimating the standard errors of the β-estimates. If we instead use the standard normal distribution, we will over-reject when we use Wald tests.
  - With cluster-robust standard errors, it turns out that using the t-distribution will not be enough to fix the standard errors. Even when bias-correcting the cluster-robust standard errors and using a t-distribution with $G-1$ degrees of freedom for Wald tests, we will with few clusters over-reject the null hypothesis of $\beta = 0$.
  - Nevertheless, it is still better than using the critical values from the t-distribution than from the standard normal distribution when we have few clusters.

3.3 What to cluster over?

It is not always obvious what to cluster over. In the example above with randomization on class level, one could fear that there is a school-level component in the errors such that clustering on school level rather than class-level is called for. However, as a practical rule of thumb one should cluster at the level of experimentation.

[8] This is easiest to intuitively understand if we consider the case of equal cluster size, that is $M_g = M$. The variance estimator in equation (**??**) was also first derived by White (1984) for the case of balanced clusters and then subsequently by Liang and Zeger for the unbalanced case.

- Clearly, we should only define clusters broad enough to allow for the error correlation we expect, but not define the groups too broad such that we just let zero correlated observations into a cluster.
- Furthermore, we should not define clusters so large that there are too few clusters.
- Sometimes aggregate explanatory variables can guide us such that we at least cluster over the same level as the aggregate variable is defined by. For example, Browning, Gørtz and Leth-Petersen (2013) study the effect of housing prices on consumption. Since all houses are not sold every year, they cannot use the individual house price and instead they use the average housing price at municipality level to estimate the effect of housing prices on consumption. Since the average housing price on municipality level is an aggregated variable, they cluster on municipality level.
- Another rule of thumb is to progressively increase the cluster size and stop whenever there is relatively little change in the estimated standard errors.
- It is sometimes the case that clusters are non-nested, for example, industries and municipalities. In this case, it will be wrong to cluster over the intersection, i.e. a particular industry in a particular municipality should not be a cluster. Instead, we should allow for $E\left(u_i u_j | x_i, x_j\right) \neq 0$ whenever individuals i and j are in the same cluster either in terms of industry or in terms of municipality.[9]

4 Bootstrapping

Bootstrapping is an easy way to obtain standard errors, confidence intervals or critical values for test statistics and p-values, but why and when it works is much more difficult.

There are two ways of obtaining standard errors for an estimate:

- Asymptotics.
- Bootstrapping.

Bootstrapping is a method for estimating the distribution of an estimator or test statistic by re-sampling one's data, that is treating the data (the sample) as if it was the population.

4.1 Why the bootstrap?

- Sometimes it is difficult to derive asymptotic distribution of an estimator or statistic, e.g. two-step estimators.

[9] The reader is referred to Cameron, Gelbach and Miller (2011) and Thompson (2011) for the theory of two-way clustering.

- Asymptotic results may be very inaccurate in finite samples and it is very difficult to derive small sample properties of estimators.
- Bootstrap approximates the distribution of an estimator or test statistic well and often more accurate in finite samples.

4.2 Why not the bootstrap?

- Computational expensive - model has to be estimated many times.
- The numerical performance of the bootstrap may be poor, when estimators whose asymptotic covariance matrices are "nearly singular", as with instrumental variable estimation with many weak instruments.
- Bootstrap is sometimes biased and should not be used blindly or uncritically.

4.3 Why is the method called the bootstrap?

Baron Münchausen is both a historical and literary character. The historical character lived in Germany in the 18th century. When he was young, he joined the Russian army where he became captain. He retired at the age of 30 and lived the rest of his life at his manor. He is known to have told witty and exaggerated stories mainly about his time in the Russian army. In one of these stories, he falls into a swamp and cannot get up. According to the story he pulled himself up by the bootstrap.[10] The point of this story is that he got himself out of the problem by using existing resources. The bootstrap uses existing data to generate the (unknown) population distribution.

4.4 The nonparametric bootstrap

There exist different types of bootstrap estimators. We will begin by considering the simplest, the nonparametric bootstrap (sometimes also called the pairs bootstrap). The nonparametric bootstrap is a very general bootstrap, which can be applied to a wide range of models including non-linear models. However, other bootstrap methods generally provide a better approximation than the nonparametric bootstrap. Furthermore, we should notice that with the nonparametric bootstrap we assume that there is no cluster-correlation. Therefore, we will briefly consider what to do in case of clustered data.

The algorithm for the nonparametric bootstrap is:

1. Estimate model on original sample to obtain the statistic T_N
2. Draw with replacement pairs $(y_i, \mathbf{x}_i)$ until a bootstrap sample of size N is reached, that is $(y_1^*, \mathbf{x}_1^*), (y_2^*, \mathbf{x}_2^*), ... (y_N^*, \mathbf{x}_N^*)$.

[10] There actually seems to be disagreement whether he pulled himself up by the bootstrap or the hair.

3. Use the bootstrap sample to obtain an estimate $T^*_{N,b}$.

4. Repeat 2)-3) many times to obtain a sequence of bootstrap estimates, $T^*_{N,b}$, $b = 1, .., B$.

5. Calculate for example the standard deviation of the B values of $T^*_{N,b}$.

Statistics

- The sample variance (from which we can compute the standard errors)

$$\frac{1}{B-1}\sum_{b=1}^{B}\left[T^*_{N,b} - \bar{T}^*_N\right]\left[T^*_{N,b} - \bar{T}^*_N\right]'$$

  where $\bar{T}^*_N$ is the mean of the B bootstrap statistics $T^*_{N,b}$.

- 95% confidence intervals can be obtain by finding 0.025 and 0.975 percentiles in the bootstrap distribution of T^*_N. If we use the percentiles in the tails of the distribution, we would need more bootstrap replications than when computing the variance.

4.5 Consistency of the bootstrap

We use $\{X_i : 1, ..., N\}$ as notation for the data, where X_i typically is a vector $(y_i, \mathbf{x}_i)$. The data are assumed to be iid draws from the population cdf $F = \Pr(X \leq x)$.

The statistic of interest is a function of $X_1, ..., X_N$ and is denoted $T_N = T_N(X_1, ..., X_N)$. This statistic has its own exact finite-sample distribution, which depends on N. Denote the finite sample distribution by $G_N = \Pr(T_N \leq t) = G_N(t, F)$. Ideally, the best thing to do would be to use this exact finite sample distribution for inference, but this is in general infeasible. Hence, for inference the problem is always to find a good approximation to G_N. When applying conventional asymptotic theory we let $N \to \infty$ and use the asymptotic distribution $G_\infty = G_\infty(t, F)$. In other words,

$$T_N \xrightarrow{d} T_\infty$$

The approach of bootstrap is quite different from this. Instead of replacing G_N with G_∞, bootstrapping replaces the population cdf F by a consistent estimator $\hat{F}$ of F. In other words, we use the bootstrap to approximate $G_N\left(\cdot, \hat{F}\right)$. There are several ways of constructing $\hat{F}$. For the nonparametric bootstrap $\hat{F}$ is the empirical distribution function

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(X_i \leq x)$$

where the function $1(\cdot)$ is an indicator function. An alternative is the parametric bootstrap, where $\hat{F}(x) = F_{\hat{\theta}}(x)$, where $F(\cdot) = F_{\theta}(\cdot)$. Hence, if $F(\cdot) = N(\mu, \sigma^2)$ we can use $\hat{F}(x) = N(\hat{\mu}, \hat{\sigma}^2)$.

The bootstrapped statistic is

$$T_N^* = T_N(X_1^*, ..., X_N^*)$$

where T_N^* is a random variable and where $X_1^*, ..., X_N^*$ are drawn randomly with replacement. The empirical cumulative distribution function of the bootstrapped statistics $T_{N,1}^*, ..., T_{N,B}^*$ is

$$\hat{G}_N\left(t, \hat{F}\right) = P\left(T_{N,b}^* \leq t\right) \tag{15}$$

Consistency of the bootstrap implies that $\left\|G_N\left(t, \hat{F}\right) - G_\infty(t, F)\right\| \xrightarrow{p} 0$ as $N \to \infty$ uniformly over all t. To show that this is the case we can split the l.h.s. in two

$$\left\|G_N\left(t, \hat{F}\right) - G_\infty(t, F)\right\| \leq \left\|G_N\left(t, \hat{F}\right) - G_N(t, F)\right\| + \|G_N(t, F) - G_\infty(t, F)\| \tag{16}$$

- Consider the first part of the r.h.s. Since $\hat{F}$ is a random variable, $G_N\left(t, \hat{F}\right)$ is also a random variable and we must refer to some kind of probabilistic convergence

$$\begin{aligned} \left\|G_N\left(t, \hat{F}\right) - G_N(t, F)\right\| &= \|P^*(T_N^* \leq t) - P(T_N \leq t)\| \\ &= \sup_t |P^*(T_N^* \leq t) - P(T_N \leq t)| \xrightarrow{p} 0 \end{aligned}$$

  where we use the Kolmogorov-Smirnov distance, which is the sup-norm. When $\hat{F}(x) \to F(x)$ and when $G_N(t, F)$ is continuous in F, $G_N\left(t, \hat{F}\right) \longrightarrow G_N(t, F)$.

- The last part deals with the uniform convergence in distribution of T_N to T_∞. If the limiting distribution $G_\infty(t, F)$ is continuous, Polya's theorem says that $G_N(t, F) \to G_\infty(t, F)$ uniformly for all t.

Therefore, we have that $\left\|G_N\left(t, \hat{F}\right) - G_\infty(t, F)\right\| \xrightarrow{p} 0$ as $N \to \infty$ uniformly over all t.

It is very important that $G_\infty(t, F)$ is continuous, otherwise the bootstrap may not work. For example, the bootstrap is inconsistent for Manski's maximum score estimator, which is a binary least absolute deviation estimator

$$\min_\beta \sum_{i=1}^N |y_i - 1(x_i\beta)|$$

However, it turns out that for median (and also quantile) regression

$$\min_{\beta} \sum_{i=1}^{N} |y_i - x_i\beta|$$

the bootstrap works.

4.6 How many bootstraps?

Obviously, using more bootstrap replications B gives a more accurate bootstrapped statistic, but in practice this must be weighted against the computational burden associated with the more replications. How to choose how many bootstrap replications to use? There exist application-specific methods for selecting the number of bootstrap replications B.[11] Cameron and Miller (2015) write that using *"*$B = 400$ *should be more than adequate in most settings".*

4.7 The residual bootstrap and the Wild bootstrap

The nonparametric bootstrap can be applied to a wide range of models, but other bootstrap methods have better properties as they assume a particular structure. For the cross-sectional linear model with additive iid errors, we can use the residual bootstrap:

1. Estimate the original model $y_i = \mathbf{x}_i\beta + u_i$ by OLS and calculate the residuals $\hat{u}_i = y_i - \mathbf{x}_i\hat{\beta}$. With N observations, we will have $\hat{u}_1, \hat{u}_2, ..., \hat{u}_N$ residuals.

2. Draw with replacement from the residuals $\hat{u}_1, \hat{u}_2, ..., \hat{u}_N$ to obtain N bootstrap residuals $u_1^*, u_2^*, ..., u_N^*$.

3. Keep the original N $\mathbf{x}$'s and use the sequence of bootstrap residuals to create N new dependent variables, that is $y_i^* = \mathbf{x}_i\hat{\beta} + u_i^*$. The bootstrap sample is then $(y_1^*, \mathbf{x}_1), (y_2^*, \mathbf{x}_2), ..., (y_N^*, \mathbf{x}_N)$.

4. Use the bootstrap sample to estimate β_b^* as $\hat{\beta}_b^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*$.

5. Repeat 2)-4) B times to obtain a sequence of bootstrap estimates, $\hat{\beta}_b^*$, $b = 1, .., B$.

6. Calculate for example the standard deviation of the B values of $\hat{\beta}_b^*$

By essentially reshuffling the residuals, we effectively break any relationship between x and u. In contrast to this, we have that under heteroscedasticity, the variance of the error term depends on x, $E\left(u_i^2|x_i\right) = \sigma_i^2$. Therefore, the so-called Wild bootstrap replaces the second step above by constructing $u_1^*, u_2^*, ..., u_N^*$ in

[11] We refer to Andrews and Buchinsky (2000) and Poi (2004).

a way which preserves the relationship between x and u. Suppose we take each residual and in 50 percent of the cases multiplied it by -1

$$u_i^* = \begin{cases} -\hat{u}_i & \text{with probability } p = 0.5 \\ \hat{u}_i & \text{with probability } 1-p \end{cases}$$

we will have that for the individual observation $(u_i^*)^2 = \hat{u}_i^2$. Moreover, we will have that three out of the first four moments are the same for the original residuals and the constructed bootstrap residuals

$$\begin{aligned} E(u_i^*) &= 0 \\ E\left((u_i^*)^2\right) &= E(\hat{u}_i^2) \\ E\left((u_i^*)^4\right) &= E(\hat{u}_i^4) \end{aligned}$$

It has been shown that the original and the constructed residuals can at most have the same moments for three out of the four first moments. Therefore, the question is which of the three moments we want to hit. Mammen (1993) shows that if we construct the new residuals as below we can instead preserve the first three moments, but not the fourth

$$u_i^* = \begin{cases} \hat{u}_i \frac{1-\sqrt{5}}{2} & \text{with probability } p = \frac{1+\sqrt{5}}{2\sqrt{5}} \\ \hat{u}_i \frac{1+\sqrt{5}}{2} & \text{with probability } 1-p \end{cases}$$

4.8 Clustering and the bootstrap

The bootstrap methods, we have dealt with, assume that errors are drawn independently. This assumption is violated with time-series data, clustered data and panel data. The bootstrap procedure for clustered data and panel data is called the block bootstrap:

1. Estimate model on original cluster sample to obtain T_N.
2. The bootstrap sample is created by repeatedly drawing a cluster (with replacement) from the sample of the G clusters and including all M_g observations for the drawn cluster, where the draws continue until the sample size $N = \sum_{g=1}^{G} M_g$ is reached.
3. Use the bootstrap sample to obtain an estimate $T_{N,b}^*$.
4. Repeat 2)-3) B times to obtain a sequence of bootstrap estimates, $T_{N,b}^*$, $b = 1, .., B$.
5. Calculate for example the standard deviation of the B values of $T_{N,b}^*$.

4.9 Examining the bootstrapped values

It is a good idea to examine the bootstrapped values to make sure that the bootstrapped standard errors are not corrupted. This is, in particular, important when using the clustered bootstrap with few clusters. Therefore, depict the bootstrapped values using a histogram (or a kernel density).

- In cases with little variation in a dummy variable (such as an treatment indicator), some bootstrap samples will have very little variation and this can give rise to very small standard errors.
- In cases with fairly high correlation between explanatory variables, some bootstrap samples will almost have multicollinearity and this can blow up the bootstrapped standard errors.
- If the parameter estimates are sensitive to the inclusion of one cluster, the histogram will show sizable probability mass for outlier estimates.

References

[1] Andrews, D.W.K. and M. Buchinsky (2000), "A Three-Step Method for Choosing the Number of Bootstrap Repetitions", *Econometrica*, Vol. 68, No. 1, pp. 23–51.

[2] Angrist, J.D., J.-S. Pischke (2009), *"Mostly Harmless Econometrics: An Empiricist's Companion"*, Princeton University Press.

[3] Bertrand, Duflo and Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?", *Quarterly Journal of Economics*, vol. 119, pp. 249-275.

[4] Browning, M., M. Gørtz and S. Leth-Petersen (2013), "Housing Wealth and Consumption: A Micro Panel Study", *Economic Journal*, Vol. 123, pp. 401-428.

[5] Cameron, A.C. and D.L. Miller (2015), "A Practitioner's Guide to Cluster-Robust Inference", *Journal of Human Resources*, Vol. 50, No. 2, pp.317-373.

[6] Chesher, A. and I. Jewitt (1987), "The Bias of a Heteroscedasticity Consistent Covariance Matrix Estimator", *Econometrica*, Vol. 55, No. 5, pp. 1217-1222.

[7] Efron, B (1979), "Bootstrap Methods: Another Look at the Jackknife", *Annals of Statistics*, Vol. 7, pp. 1-26.

[8] Horowitz, J.L. (2001), "The Bootstrap", in J.J. Heckman and E. Leamer (eds.): *"Handbook of Econometrics"*, Vol. 5, Elsevier Science.

[9] MacKinnon, J.G., and H. White (1985), "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties", *Journal of Econometrics*, Vol. 29, pp. 305-325.

[10] Mammen, E. (1993), "Bootstrap and Wild Bootstrap for High-dimensional Linear Models", *Annals of Statistics*, Vol. 21, pp. 255-285.

[11] Moulton, B.R. (1986), "Random Group Effects and the Precision of Regression Estimates", *Journal of Econometrics*, Vol. 32, No. 2, pp. 385-397.

[12] Poi, B.P. (2004), "From the Help Desk: Some Bootstrapping Techniques", *Stata Journal*, Vol. 4, No. 3, pp. 312-328.

[13] White, H. (1980a), "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity", *Econometrica*, Vol. 48, No. 4, pp. 817-838.

[14] White, H. (1980b), "Using Least Squares to Approximate Unknown Regression Functions", *International Economic Review*, Vol. 21, No. 1, pp. 149-170.

[15] White, H. (1984), *"Asymptotic Theory for Econometricians"*, Academic Press, San Diego.