

Applied Econometrics

Lecture Notes

Alessandro Tarozzi*

1 Estimation with robust standard errors

Economists do not care only about point estimates. We also care about standard errors, as these allow us to measure the level of ‘confidence’ we can have on the meaningfulness of the point estimates. For this reason, getting the standard errors right is important. A lot of standard results in econometrics are based on the assumption that observations are independent, and identically distributed. But in applied microeconomics, data often come from surveys, and surveys are typically designed in such a way that observations are **not i.i.d.** Forgetting about this leads to sometimes wildly wrong inference. In particular, we will see that it frequently leads to standard errors that are too low. Notice that this problem goes beyond the ‘usual’ warning against the presence of heteroskedasticity.

Knowing the methodology of data collection is important!

1. Does the sampling frame really coincide with population of interest? (Homeless, or military, often not included in sampling frame). See Figure 1.1 for Taiwan in [Deaton \(1997\)](#).
2. Data are, almost always, **not as objective** as we might think (or hope). The way you collect data (the order of the questions in a questionnaire, or the way you phrase them) can affect importantly the kind of answers you get! If you are interested in survey methodology issues, especially as related to the collection of expenditure data, a great starting point is [Deaton and Grosh \(2000\)](#).
3. More to the point, SURVEY DESIGN should be taken into account for:
 - Sampling weights (different population units have a different probability of being selected. The use of weights typically changes the point estimates, and not only the standard errors).
 - Stratification (makes observations not identically distributed, and might change standard errors, especially in small samples)
 - Clustering (makes observations not independent, and frequently increases standard errors, even in large samples)

*I would like to thank Benjamin Zhang for his help in writing a first draft of some chapters of these lecture notes.

For many surveys it is essential to correct S.E.

CPS (USA), Panel study of Income Dynamics (USA), Survey of Income & Program Participation (USA), Health & Retirement Study (USA), National Sample Survey (India), Living Standard Measurement Surveys (World Bank surveys), Demographic and Health Surveys (DHS).

An excellent introduction to Stratification/Clustering/Sampling weights can be found in Ch. 1 of [Deaton \(1997\)](#).

1.1 Stratification

First split the survey into more INDEPENDENT sub-surveys, and *then* draw independent samples from each stratum. This is typically done to ensure that enough observations come from different areas / groups. Therefore, by construction, observations from different strata belong to at least partly different populations! This means that they are **not** identically distributed.

Stratification typically **reduces** SE especially in small samples (and especially if parameters of interest are very different across different strata. Why? Intuitively, because all possible samples become ‘more similar’).

1.2 Clustering

This can coexist with stratification. Sampling is done using a two-stage design:

1. First select **clusters/PSU** (Primary Stage Units). These are typically urban blocks, villages, neighborhoods etc.
2. then select households (or individuals).

In most cases clustering **increases** (sometimes considerably) standard errors. Why? Intuitively, because observations within the same cluster are generally correlated among them, and this reduces the ‘effective number of observations’ (we will see that if the intra-cluster correlation goes up, SE tend to go up). So why is it done? Often to keep survey costs at bay (not necessary, for example, if you use a telephone survey!). Also, if you are doing a study based on a randomized evaluation, and treatments and controls are defined at the group level (e.g. by village), then clustering arises by construction.

Correcting SE for clustering mean, essentially, allowing for intra-group correlation. Intra-group correlation can also of course arise for reasons different from survey design (e.g. you have more observations per unit, as in a panel).

[Kish \(1965\)](#) Deff (Design Effect Ratio)

$$\frac{Var(\text{Taking into account design})}{Var(\text{Assuming SRS})} \quad \text{typically } > 1$$

1.3 Sampling Weights

There are also called ‘inflation factors’, as each observations is ‘inflated’ to represent a certain share of the population. Broadly speaking

$$Weight = \frac{1}{\text{Prob. of Selection}} \quad \text{‘Inflation Factor’}$$

Different hhs/individuals typically have \neq probability of being selected. This can happen for different reasons:

1. Because of two stage sampling: for ex., urban blocks (PSU) are selected with the same prob., and then a constant number of hhs is selected in each block. Because blocks have different numbers of hhs, hhs in larger blocks have **lower** probability of being selected. Stratification and clustering are generally associated with the presence of sampling weights, as they typically lead to probability of selection that differs across population units. Some surveys, however, are self weighted (because the number of observations selected in each first stage unit is proportional to the ‘size’ of the unit) but in practice post-sampling weights is frequently necessary to adjust for non-responses (these ‘adjustments’ have frequently some degree of subjectivity...).
2. To take into account \neq cost of sampling in \neq areas, or \neq non-response rates.
3. To increase precision (high prob. of selection for units that for ex., contribute more to the mean. The ‘contribution’ is generally determined using proxies).

Here we will focus on weights deriving from (1), which is the case that Wooldridge calls **Standard Stratified Sampling** (see [Wooldridge \(2002\)](#) (W) 17.8). There are many other (generally more complicated) cases where ‘reweighting’ is necessary because of missing data/non-response (see W 17.7).

Standard Stratified Sampling Consider the case in which the population is divided into S groups or strata, and a sample of n_s observations is selected from each stratum (and $\sum_{j=1}^S n_s = n$). If the strata are not homogeneous, this kind of sampling BREAKS identical distribution (which, however, *remains* within each stratum). Broadly speaking, if strata are different, and you don’t use weights, you can’t estimate consistently populations parameters. Let $Q_j \equiv$ **population** frequency of stratum j . You want to estimate the parameters θ_0 that satisfy a certain *population* moment condition:

$$E[q(x_{ij}, \theta_o)] = 0 \quad \text{pop. moment condition}$$

where x_i are variables observed for individual i . Using the L.I.E., we easily get

$$\sum_j^S Q_j E[q(x_{ij}; \theta_o) | j] = 0$$

where I have made explicit the fact that each unit i belongs to one and only one stratum. Let $H_j \equiv$ fraction of **sample** from stratum j . Note that H_j is defined by the sampling scheme, and it remains *approximately constant when the sample size increases!*

Suppose now that you estimate θ_0 by using an unweighted sample analogue of the above population moment condition. You have a sample of n observations, $H_j = n_j/n$ of which come from stratum s . So

$$\begin{aligned} \frac{1}{n} \sum_{h=1}^n q(x_h; \theta_o) &= \frac{1}{n} \sum_{j=1}^S \sum_{i=1}^{n_j} q(x_{ij}; \theta_o) = \frac{1}{n} \sum_{j=1}^S \sum_{i=1}^{n_j} \frac{n_j}{n_j} q(x_{ij}; \theta_o) \\ &= \sum_{j=1}^S H_j \left[\frac{1}{n_j} \sum_{i=1}^{n_j} q(x_{ij}; \theta_o) \right] \end{aligned}$$

When the sample size increases, we get more and more observations from each stratum, but the number of strata does not change, and the proportion of observations from each stratum does not change as well. So, under regularity conditions, the object in brackets will converge in probability to $E[q(x_{ij}; \theta_o) | j]$, but unless $H_j = Q_j$ (which typically does not happen!), the sample quantity $\frac{1}{n} \sum_{h=1}^n q(x_h)$ does **not** identify the parameter of interest θ_o !!

But this is no big deal. We just have to incorporate sampling weights in our estimation, and use instead

$$\frac{1}{n} \sum_{j=1}^S \sum_{i=1}^{n_j} \frac{Q_j}{H_j} q(x_{ij}; \theta_o)$$

where $\frac{Q_j}{H_j} \equiv W_{ji}$ is the SAMPLING WEIGHT. In fact, now

$$\frac{1}{n} \sum_{j=1}^S \sum_{i=1}^{n_j} \frac{Q_j}{H_j} q(x_{ij}; \theta_o) = \sum_{j=1}^S Q_j \left[\frac{1}{n_j} \sum_{i=1}^{n_j} q(x_{ij}; \theta_o) \right]$$

so that now the sample moment condition **does** identify the parameter of interest θ_o , as the sample moment condition will ‘converge’ to the correct population moment condition.

1.4 Variance of sample mean, with stratification.

A mixture of Deaton (1.38) & Wooldridge (Ch. 17.8)

Let us see, formally, that stratification typically reduces standard errors. For simplicity I consider the case in which the proportion of the population living in each stratum is known. Here, once again, denote Q_j prop. of population in stratum j , and H_j prop. of sample from stratum j . The sampling weights are typically identical for all observations in the same stratum:

$$\begin{aligned} w_{jl} &= \frac{Q_j}{H_j} = \frac{Q_j}{n_j/n} \implies \sum_{j=1}^J \sum_{l=1}^{n_j} w_{jl} = \sum_j^J n_j \frac{Q_j}{n_j/n} = n \\ &\implies \frac{w_{jl}}{\sum_{j=1}^J \sum_{l=1}^{n_j} w_{jl}} = \frac{Q_j}{n_j} \end{aligned}$$

First note that the sample mean can be written as a weighted average of stratum-specific sample means.

$$\begin{aligned}\bar{x} &= \sum_{j=1}^J \sum_{l=1}^{n_j} \left(\frac{w_{jl}}{\sum_{j=1}^J \sum_{l=1}^{n_j} w_{jl}} \right) x_{jl} = \sum_{j=1}^J \sum_{l=1}^{n_j} \left(\frac{Q_j}{n_j} \right) x_{jl} \\ &= \sum_{j=1}^J Q_j \left[\frac{1}{n_j} \sum_{l=1}^{n_j} x_{jl} \right] = \sum_j Q_j \bar{x}_j\end{aligned}$$

Since strata are independent, the ‘correct’ variance of the sample average can be estimated as a weighted average of the stratum-specific estimated variances. So

$$\widehat{Var}_{STRAT}(\bar{x}) = \sum_j Q_j^2 \widehat{Var}(\bar{x}_j) = \sum_j Q_j^2 \frac{S_j^2}{n_j} = \sum_j Q_j^2 \left[\frac{1}{n_j} \frac{1}{n_j - 1} \sum_{l=1}^{n_j} (x_{jl} - \bar{x}_j)^2 \right] \quad (1)$$

where we use the fact that within each stratum weights play no role, so that the stratum-specific variance can be simply estimated using

$$S_j^2 = \frac{1}{n_j - 1} \sum_{l=1}^{n_j} (x_{jl} - \bar{x}_j)^2$$

What is the relation between **1** and the variance as it would be computed disregarding the presence of strata? First of all, let us compute this latter variance. Even if stratification is not considered, one still has to make use of the correct formula for the variance when sampling weights are present (which is the case here). Using expression (1.28) in Deaton (and abstracting from the finite population correction $n/(n-1)$),

$$\widehat{Var}_{SRS}(\bar{x}) = \sum_j \sum_{l=1}^{n_j} \left(\frac{Q_j}{n_j} \right)^2 (x_{jl} - \bar{x})^2$$

How do $\widehat{Var}_{SRS}(\bar{x})$ and $\widehat{Var}_{STRAT}(\bar{x})$ differ? Let us rewrite $\widehat{Var}_{STRAT}(\bar{x})$:

$$\begin{aligned}\widehat{Var}_{STRAT}(\bar{x}) &= \sum_j Q_j^2 \frac{1}{n_j - 1} \frac{1}{n_j} \sum_{l=1}^{n_j} (x_{jl} - \bar{x}_j \pm \bar{x})^2 \\ &= \sum_j Q_j^2 \frac{1}{n_j - 1} \frac{1}{n_j} \sum_{l=1}^{n_j} [(x_{jl} - \bar{x}) - (\bar{x}_j - \bar{x})]^2 \\ &= \sum_j Q_j^2 \frac{1}{n_j - 1} \frac{1}{n_j} \left[\sum_{l=1}^{n_j} (x_{jl} - \bar{x})^2 + \sum_{l=1}^{n_j} (\bar{x}_j - \bar{x})^2 - 2 \sum_{l=1}^{n_j} (x_{jl} - \bar{x})(\bar{x}_j - \bar{x}) \right] \\ &= \sum_j Q_j^2 \frac{1}{n_j - 1} \frac{1}{n_j} \left[\sum_{l=1}^{n_j} (x_{jl} - \bar{x})^2 + n_j (\bar{x}_j - \bar{x})^2 - 2 (\bar{x}_j - \bar{x}) \sum_{l=1}^{n_j} (x_{jl} - \bar{x}) \right] \\ &= \sum_j Q_j^2 \frac{1}{n_j - 1} \frac{1}{n_j} \left[\sum_{l=1}^{n_j} (x_{jl} - \bar{x})^2 - n_j (\bar{x}_j - \bar{x})^2 \right]\end{aligned}$$

so that the following approximation will hold:

$$\begin{aligned}\widehat{Var}_{STRAT}(\bar{x}) &\approx \sum_j^J \sum_{l=1}^{n_j} \left(\frac{Q_j}{n_j}\right)^2 (x_{jl} - \bar{x})^2 - \left[\sum_j^J Q_j^2 \frac{1}{n_j - 1} (\bar{x}_j - \bar{x})^2 \right] \\ &= \widehat{Var}_{SRS}(\bar{x}) - \sum_j^J Q_j^2 \frac{1}{n_j - 1} (\bar{x}_j - \bar{x})^2,\end{aligned}$$

which means that taking stratification into account decreases the variance of the estimator. Note that the difference in the variances is larger the larger the cross-strata differences in means are. Note also that for large samples the difference is unlikely to be large, since it converges to zero when the sample size from each stratum grows. In practice, taking stratification into account often makes little important difference.

2 A unifying Method of Moments approach

We assume that there is only **one stratum**.¹ Here we loosely develop arguments leading to the asymptotic variance of the MoM estimator when the number of clusters goes to infinity. This framework is analogous to the ‘large-N’ asymptotic results for panel estimation. Let n be the number of clusters, and m_i is the number of observations within cluster i . The estimator solves the following sample moment condition (for simplicity we consider the case of exact identification, with k moments and k parameters):

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^{m_i} w_i q(x_{ji}; \hat{\theta}) &= 0 \\ \sum_{i=1}^n g_i(\hat{\theta}) &= 0\end{aligned}$$

Where $g_i(\hat{\theta}) \equiv \sum_{j=1}^{m_i} w_i q(x_{ji}; \hat{\theta})$ (this is basically the weighted sum of the ‘errors’ within the same cluster). Note that in the above expression the sampling weight w_i is just a scalar. Expanding $q(x_{ji}; \hat{\theta})$ around θ_0 we get

$$q(x_{ji}; \hat{\theta}) = q(x_{ji}; \theta_0) + \frac{\partial q(x_{ji}; \tilde{\theta})}{\partial \theta'} (\hat{\theta} - \theta_0)$$

¹For a more formal and general treatment of GMM estimation with a complex survey design see [Bhattacharya \(2005\)](#).

Plugging into the moment condition and rearranging (here we abstract from regularity conditions, see for example W, page 351), we get

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta_0) &= \left[-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} w_i \frac{\partial q(x_{ji}; \tilde{\theta})}{\partial \theta'} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\sum_{j=1}^{m_i} w_i q(x_{ji}; \theta_0) \right] \\ &= \left[-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{m_i} w_i \frac{\partial q(x_{ji}; \tilde{\theta})}{\partial \theta'} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g_i(\theta_0) \\ &\xrightarrow{d} N(0, A_0^{-1} B_0 A_0^{-1})\end{aligned}$$

where

$$\begin{aligned}A_0 &= E \left[\frac{\partial q(x_{ji}; \theta_0)}{\partial \theta'} \right] \\ B_0 &= E [g_i(\theta_0) g_i(\theta_0)'] = E \left[w_i^2 \left(\sum_{j=1}^{m_i} q(x_{ji}; \theta_0) \right) \left(\sum_{j=1}^{m_i} q(x_{ji}; \theta_0) \right)' \right]\end{aligned}$$

so note that **the presence of clustering will not affect the estimation of the Hessian**, but it WILL affect the estimation of the variance-covariance matrix through B_0 . Consistent estimates of the variance can therefore be obtained as

$$\begin{aligned}\frac{1}{n} \hat{A}_0^{-1} \hat{B}_0 \hat{A}_0^{-1} \text{ where} \\ \hat{A}_0 &= \frac{1}{n} \sum_i \sum_{j=1}^{m_i} w_i \frac{\partial q(x_{ji}; \hat{\theta})}{\partial \theta'} \\ \hat{B}_0 &= \frac{1}{n} \sum_i w_i^2 \left(\sum_{j=1}^{m_i} q(x_{ji}; \hat{\theta}) \right) \left(\sum_{j=1}^{m_i} q(x_{ji}; \hat{\theta}) \right)'\end{aligned}$$

2.1 Asymptotic variance for the sample mean, with sampling weights but no clustering

Since here we have no clustering, n will be the number of observations, and the index j is irrelevant. For a sample mean, $q(x_{ji}; \hat{\theta}) = x_i - \bar{x}$, so that

$$\begin{aligned}\hat{B}_0 &= \frac{1}{n} \sum_i w_i^2 (x_i - \bar{x})^2 \\ \hat{A}_0 &= -\frac{1}{n} \sum_i w_i\end{aligned}$$

so that one can estimate the variance of the weighted mean using

$$\begin{aligned}
\widehat{Var}(\bar{x}) &= \frac{1}{n} \hat{A}_0^{-1} \hat{B}_0 \hat{A}_0^{-1} \\
&= \left(\sum_i^n w_i \right)^{-1} \sum_i^n w_i^2 (x_i - \bar{x})^2 \left(\sum_i^n w_i \right)^{-1} \\
&= \sum_i^n \left(\frac{w_i}{\sum_i^n w_i} \right)^2 (x_i - \bar{x})^2
\end{aligned} \tag{2}$$

which is identical to (1.28) in Deaton (1997) up to the (negligible for n large) fpc $n/(n-1)$. Note that this is different from the ‘standard’ form of the variance

$$\widehat{Var}(\bar{x}) \neq \frac{S^2}{n} = \frac{1}{n} \frac{1}{n-1} \sum_i^n w_i (x_i - \bar{x})^2 \tag{3}$$

2.2 Asymptotic variance for the sample mean, with sampling weights and clustering

The estimator for A_0 is not affected by clustering. In fact

$$\hat{A}_0 = \frac{1}{n} \sum_i^n \sum_{j=1}^{m_i} w_i = \frac{1}{n} \sum_i^n w_i m_i$$

The estimate of B_0 , instead, is affected.

$$\begin{aligned}
B_0 &= E \left[w_i^2 \left(\sum_{j=1}^{m_i} (x_{ji} - \mu_0) \right) \sum_{j=1}^{m_i} (x_{ji} - \mu_0) \right] \\
&= E \left[w_i^2 (m_i \bar{x}_i - m_i \mu_0) (m_i \bar{x}_i - m_i \mu_0) \right] = E \left[w_i^2 m_i^2 (\bar{x}_i - \mu_0)^2 \right]
\end{aligned}$$

which can be consistently estimated using

$$\hat{B}_0 = \frac{1}{n} \sum_i^n w_i^2 m_i^2 (\bar{x}_i - \bar{x})^2$$

so that

$$Var(\bar{x}) \stackrel{a}{\approx} \frac{1}{n} \hat{A}_0^{-1} \hat{B}_0 \hat{A}_0^{-1} = \sum_i^n \left(\frac{w_i m_i}{\sum_i^n w_i m_i} \right)^2 (\bar{x}_i - \bar{x})^2$$

which is the same as (1.52) in Deaton (97) up to the ‘usual’ fpc.

2.3 Asymptotic variance for OLS, with clustering and sampling weights

The formulae developed here can also be used when sampling weights are not present, but one wants to adjust for the presence of intra-group correlation. In the case of OLS, for example,

$$\begin{aligned}
 q(x_{ji}; \hat{\theta}) &= \underset{(k \times 1)}{\mathbf{x}_{ij}} \left(\underset{(1 \times 1)}{y_{ij}} - \underset{(1 \times k)}{\mathbf{x}'_{ij}} \underset{(k \times 1)}{\hat{\beta}} \right) = \underset{(k \times 1)}{\mathbf{x}_{ij}} e_{ij} \\
 \frac{\partial q(x_{ji}; \hat{\theta})}{\partial \theta'} &= \underset{(k \times k)}{-\mathbf{x}_{ij} \mathbf{x}'_{ij}} \\
 \hat{A}_0 &= -\frac{1}{n} \sum_i^n \sum_{j=1}^{m_i} w_i \mathbf{x}_{ij} \mathbf{x}'_{ij} = -\frac{1}{n} \sum_i^n w_i \mathbf{X}'_i \mathbf{X}_i
 \end{aligned}$$

where

$$\underset{(m_i \times k)}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{im_i} \end{bmatrix}$$

and

$$\begin{aligned}
 g_i(\hat{\beta}) &= \sum_{j=1}^{m_i} w_i \mathbf{x}_{ij} \left(\underset{(k \times 1)}{y_{ij}} - \mathbf{x}'_{ij} \hat{\beta} \right) \\
 \hat{B}_0 &= \frac{1}{n} \sum_i^n w_i^2 \left[\sum_{j=1}^{m_i} \mathbf{x}_{ij} e_{ij} \right] \left[\sum_{j=1}^{m_i} \mathbf{x}_{ij} e_{ij} \right]' = \frac{1}{n} \sum_i^n w_i^2 \mathbf{X}'_i e_i e_i' \mathbf{X}_i
 \end{aligned}$$

where²

$$\underset{(m_i \times 1)}{e_i} = \begin{bmatrix} y_{i1} - \mathbf{x}'_{i1} \hat{\beta} \\ \vdots \\ y_{im_i} - \mathbf{x}'_{im_i} \hat{\beta} \end{bmatrix}$$

so that

$$\widehat{Var}(\hat{\beta}_{OLS}) \approx \left(\sum_i^n w_i \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \sum_i^n w_i^2 \mathbf{X}'_i e_i e_i' \mathbf{X}_i \left(\sum_i^n w_i \mathbf{X}'_i \mathbf{X}_i \right)^{-1}$$

which is (2.30) in Deaton (1997). With no clustering (so that now there is no index j and the index i denotes a single observation and \mathbf{X}_i is the vector of regressors for individual i) the formula is easily adapted to the case in which there is no serial correlation but we want to allow for heteroskedasticity. The result (in which n is now the number of observations) is the standard White heteroskedasticity-robust variance (2.33 in Deaton (1997)).

$$\widehat{Var}(\hat{\beta}_{OLS}) \approx \left(\sum_i^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \sum_i^n e_i^2 \mathbf{X}_i \mathbf{X}'_i \left(\sum_i^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1},$$

² In fact $\mathbf{X}'_i e_i e_i' \mathbf{X}_i = \begin{bmatrix} \mathbf{x}_{i1} & \cdots & \mathbf{x}_{im_i} \end{bmatrix} \begin{bmatrix} e_{i1} \\ \vdots \\ e_{im_i} \end{bmatrix} \begin{bmatrix} e_{i1} & \cdots & e_{im_i} \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{im_i} \end{bmatrix}$

Suppose now that there is homoskedasticity, that observations are independent across clusters, and that they are equi-correlated within cluster. So $\text{var}(e_{ij}) = \sigma^2 \forall i, j$, $\text{cov}(e_{ij}, e_{ih}) = \rho\sigma^2$ (same cluster, different individuals), and $\text{cov}(e_{ij}, e_{hp}) = 0$ if $i \neq h$ (different clusters, different individuals).

$$\begin{aligned}
E \left[\sum_i^n \mathbf{X}'_i e_i e'_i \mathbf{X}_i \mid \mathbf{X}_1 \cdots \mathbf{X}_n \right] &= \sigma^2 \sum_i^n \begin{bmatrix} \mathbf{x}_{i1} & \cdots & \mathbf{x}_{im_i} \end{bmatrix} \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \cdots & \rho & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{im_i} \end{bmatrix} \\
&= \sigma^2 \sum_i^n \begin{bmatrix} \rho \sum_j \mathbf{x}_{ij} + (1 - \rho) \mathbf{x}_{i1} & \cdots & \rho \sum_j \mathbf{x}_{ij} + (1 - \rho) \mathbf{x}_{im_i} \end{bmatrix} \begin{bmatrix} \mathbf{x}'_{i1} \\ \vdots \\ \mathbf{x}'_{im_i} \end{bmatrix} \\
&= \sigma^2 \sum_{i=1}^n \left[\rho \sum_{j=1}^{m_i} \mathbf{x}_{ij} \mathbf{x}'_{i1} + (1 - \rho) \mathbf{x}_{i1} \mathbf{x}'_{i1} + \dots + \rho \sum_{j=1}^{m_i} \mathbf{x}_{ij} \mathbf{x}'_{im_i} + (1 - \rho) \mathbf{x}_{im_i} \mathbf{x}'_{im_i} \right] \\
&= \sigma^2 \left\{ (1 - \rho) \sum_i^n \sum_{j=1}^{m_i} \mathbf{x}_{ij} \mathbf{x}'_{ij} + \rho \sum_i^n \left[\sum_{j=1}^{m_i} \mathbf{x}_{ij} \sum_{j=1}^{m_i} \mathbf{x}'_{ij} \right] \right\}.
\end{aligned}$$

Suppose for simplicity that m units are selected from each cluster. So $m_i = m_h = m$. Suppose *also* that, *within* each cluster, all units share the same covariates. This can happen, for example, if we are studying how cluster-specific variables (such as presence of hospitals or schools within the cluster, or local prices) affect behavior. Then $\mathbf{x}_{ij} = \mathbf{x}_{ih} \forall h, j$ in a given cluster i . This implies that $\sum_j \mathbf{x}_{ij} (\sum_j \mathbf{x}'_{ij}) = m \sum_j \mathbf{x}_{ij} \mathbf{x}'_{ij}$. So,

$$\begin{aligned}
E \left[\sum_i^n \mathbf{X}'_i e_i e'_i \mathbf{X}_i \mid \mathbf{X}_1 \cdots \mathbf{X}_n \right] &= \sigma^2 [(1 - \rho) + m\rho] \sum_i^n \sum_j \mathbf{x}_{ij} \mathbf{x}'_{ij} \\
&= \sigma^2 [1 + \rho(m - 1)] \sum_i^n \mathbf{X}'_i \mathbf{X}_i,
\end{aligned}$$

and the variance of $\hat{\beta}_{OLS}$ simplifies to

$$\begin{aligned}
\text{Var}(\hat{\beta}_{OLS}) &= \sigma^2 [1 + \rho(m - 1)] \left(\sum_i^n \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \\
&= [1 + \rho(m - 1)] \text{Var}_{naive}(\hat{\beta}_{OLS})
\end{aligned}$$

where $\text{Var}_{naive}(\hat{\beta}_{OLS})$ is the variance one would get *not* taking clustering into account! The true variance can be MUCH bigger than $\text{Var}_{naive}(\hat{\beta}_{OLS})$ unless $m = 1$ (which basically means no clustering, as we would be sampling one observation per cluster) or unless $\rho = 0$ (which means no intracluster correlation). But if, for example, we sample 10 units per cluster, and $\rho = 0.1$ (which is quite small, and very well possible!), then $[1 + \rho(m - 1)] \approx 2$, and the correct variance is about **twice** as large as the incorrect one we would get assuming no intra-cluster correlation!

One important case in which it is true that each individual share the same covariates is when we are calculating a sample mean, which basically correspond to the case where we simply regress a variable on a constant. So \mathbf{x}_{ij} is just a scalar equal to one, for each observation. Then

$$\begin{aligned} \text{Var}(\hat{\beta}_{OLS}) &= \text{Var}(\bar{y}) = \sigma^2 [1 + \rho(m-1)] \left(\sum_i^n \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \\ &= \sigma^2 [1 + \rho(m-1)] \left(n \begin{bmatrix} \mathbf{1} & \cdots & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \vdots \\ \mathbf{1} \end{bmatrix} \right)^{-1} \\ &= \frac{\sigma^2}{nm} [1 + \rho(m-1)] \end{aligned}$$

which, again, can be MUCH bigger than $\frac{\sigma^2}{nm}$ (the ‘standard’ variance of a sample mean) unless $m = 1$, or $\rho = 0$.

2.4 OLS, no clustering, with sampling weights

$$\begin{aligned} \hat{A}_0 &= \frac{1}{n} \sum_i^n w_i \mathbf{x}_i \mathbf{x}'_i = \frac{1}{n} \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ \underbrace{(k \times 1)} & & \underbrace{(k \times n)} \end{bmatrix} \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix} \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \\ \hat{B}_0 &= \frac{1}{n} \sum_i^n w_i^2 e_i^2 \mathbf{x}_i \mathbf{x}'_i = \frac{1}{n} \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} (w_1 e_1)^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & (w_n e_n)^2 \end{bmatrix} \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \end{aligned}$$

References

- BHATTACHARYA, D. (2005): “Asymptotic Inference from Multi-stage Samples,” *Journal of Econometrics*, 126(1), 145–171.
- DEATON, A. (1997): *The Analysis of Household Surveys: A Microeconomic Approach to Development Policy*. The Johns Hopkins University Press (for the World Bank).
- DEATON, A., AND M. GROSH (2000): “Consumption,” in *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*, ed. by M. Grosh, and P. Glewwe, vol. 1, chap. 5, pp. 91–133. Oxford University Press for the World Bank.
- KISH, L. (1965): *Survey Sampling*. John Wiley.
- WOOLDRIDGE, J. (2002): *Econometrics of Cross Section and Panel Data*. MIT Press.