Least Squares Estimation

Paolo Zacchia

Statistics and Econometrics

Lecture 8

Statistical properties of OLS: roadmap

- This lecture develops the **statistical properties** of OLS: they are necessary to conduct statistical inference on **b**.
- The benchmark is based on the **large sample properties** that are based on asymptotic results, with fewer statistical assumptions.
- Yet *some* assumptions are necessary: all six assumptions by White (1980) for i.n.i.d. data are reviewed and motivated.
- The traditional **small sample properties**, based on *exact* probabilistic results, are also reviewed.
- The final part of the lecture covers the case of **dependent observations** and related inference issues.

Linearity

Assumption 1

Linearity. The data are generated by a linear model that has a "true" parameter vector $\boldsymbol{\beta}_0$.

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i$$

- This assumption may appear obvious, but it helps rule out any "specification errors" about the functional form linking y_i with $\mathbf{x}_i$.
- It also formalizes the "true" parameter vector $\boldsymbol{\beta}_0$.
- Going forward, the OLS estimator of $\boldsymbol{\beta}_0$ is written as $\widehat{\boldsymbol{\beta}}_{OLS}$.

OLS decomposition

By linearity, the OLS estimator can be decomposed as follows.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i \left(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_i\right) \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \varepsilon_i\end{aligned}$$

This decomposition can also be written using compact notation.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \left(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \frac{1}{N}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\end{aligned}$$

This decomposition is extremely useful for the ensuing analysis.

Independent observations

Assumption 2

Independently but not identically distributed data. The observations in the sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$ are *independently*, but *not* necessarily *identically*, distributed (i.n.i.d.).

- The key assumption here is *independence*.
- Other than this, no particular distributional restriction is placed on the sample.
- Different sample observations can be definitely drawn from different distributions!

Statistical properties of the regressors

Assumption 3

Moments and realizations of the regressors. The random vector of regressors $\boldsymbol{x}_i$ has a finite second moment, and for some $\delta > 0$:

$$\mathbb{E}\left[|X_{ik}X_{i\ell}|^{1+\delta}\right] < \infty$$

for $k, \ell = 1, \ldots, K$ and $i = 1, \ldots, N$. In addition, its realizations $\mathbf{x}_i$ are such that, for any two $K \times 1$ vectors $\boldsymbol{\beta}'$ and $\boldsymbol{\beta}''$:

$$\mathbf{X}\boldsymbol{\beta}' = \mathbf{X}\boldsymbol{\beta}'' \text{ iff } \boldsymbol{\beta}' = \boldsymbol{\beta}''$$

thus, $\mathbf{X}$ has full column rank and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is nonsingular.

- This fully specifies the statistical restrictions placed on the regressors $\mathbf{X}$. First, a Ljapunov condition is imposed.
- It also imposes uniqueness of the Least Squares solution $\mathbf{b}$ (ruling out e.g. the dummy variable trap).

Consequences of random regressors

- Note that unlike classical treatments of regression, here the regressors are not assumed "fixed" in repeated samples.

- An asymptotic consequence of Assumption 3 is that:

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] \equiv \mathbf{K}_0$$

  where the $K \times K$ matrix $\mathbf{K}_0$ has full rank.

- This is just an application of the Law of Large Numbers for i.n.i.d. data.

- In the **special case** where the data are i.i.d. this matrix is written more simply as $\mathbf{K}_0 = \mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$.

Exogeneity

Assumption 4

Exogeneity. Conditional on the regressors $\boldsymbol{x}_i$, the error term ε_i has mean zero (using typical terminology, it is said to be **mean independent** of the regressors $\boldsymbol{x}_i$).

$$\mathbb{E}\left[\varepsilon_i \middle| \boldsymbol{x}_i\right] = 0$$

- The is is the key assumption that determines **consistency** of the estimator. Recall how this implies that $\mathbb{E}\left[\boldsymbol{x}_i \varepsilon_i\right] = 0$.
- Consistency is thus obtained from the OLS decomposition by making the following asymptotic observation.

$$\frac{1}{N}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i \xrightarrow{p} \mathbf{0}$$

- The term "exogeneity" comes from the traditional analysis of simultaneous equations models (see Lectures 9 and 10).

Heteroscedasticity (1/2)

Assumption 5

Heteroscedastic, Independent Errors. The variance of the error term ε_i conditional on $\mathbf{x}_i$ is left unrestricted (this property is called *heteroscedasticity*). Since observations are independent, the conditional covariance between the two error terms from two different observations $i, j = 1, \ldots, N$ is zero.

$$\mathbb{E}\left[\varepsilon_i^2 \middle| \boldsymbol{x}_i\right] = \sigma^2(\boldsymbol{x}_i) \equiv \sigma_i^2$$
$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{x}_i, \boldsymbol{x}_j\right] = 0$$

In addition, for some $\delta > 0$ the following holds:

$$\mathbb{E}\left[\left|\varepsilon_i^2\right|^{1+\delta}\right] < \infty$$

for all $i = 1, \ldots, N$. **(Continues...)**

Heteroscedasticity (2/2)

Assumption 5

(Continued.) All of the above can be written in compact matrix notation as follows.

$$\boldsymbol{\Sigma} \equiv \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

- Heteroscedasticity is a natural property of both social and economic data.
- The more restrictive **homoscedasticity** hypothesis, which postulates a constant conditional variance of the error term:

$$\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right] = \sigma_0^2 \mathbf{I}$$

must be treated as a special case.

Joint moments of the data

Assumption 6

Moments of $\boldsymbol{x}_i\varepsilon_i$. For $i = 1, \ldots, N$, matrix $\mathbb{E}\left[\varepsilon_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}\right]$ exists, is finite, semi-definite positive and it has full rank K. Furthermore, for some $\delta > 0$, for $i = 1, \ldots, N$, and for $k, \ell = 1, \ldots, K$, the following Ljapunov condition holds.

$$\mathbb{E}\left[\left|\varepsilon_i^2 X_{ik} X_{i\ell}\right|^{1+\delta}\right] < \infty$$

- This assumption allows to establish a proper Central Limit Theorem result for this statistical model.
- In what follows, it is useful to define the following limiting variance.

$$\boldsymbol{\Xi}_0 \equiv \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \boldsymbol{x}_i \varepsilon_i\right]$$

Analysis of the limiting variance

- Under Assumption 2 (*independent observations*) matrix $\boldsymbol{\Xi}_0$ assumes a more straightforward expression:

$$\boldsymbol{\Xi}_0 = \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{x}_i\varepsilon_i\right] = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i\varepsilon_i\right]$$
$$= \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$$

  and note that it is semi-definite positive and has full rank by Assumption 6.

- If the observations were also *identically distributed*, it then follows that $\boldsymbol{\Xi}_0 = \mathbb{E}\left[\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$. This is a special case.

- Note that under homoscedasticity, it is $\boldsymbol{\Xi}_0 = \sigma_0^2\mathbf{K}_0$.

$$\mathbb{E}\left[\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] = \mathbb{E}\left[\varepsilon_i^2\right]\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] = \sigma_0^2\,\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right]$$

A central limit theorem for OLS

- Assumption 6 allows to apply some Law of Large Numbers whereby:

$$\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \xrightarrow{p} \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] = \boldsymbol{\Xi}_0$$

- . . . and by the Ljapunov condition specified by Assumption 6, the following Central Limit Theorem result holds too.

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Xi}_0\right)$$

- This allows to finally establish asymptotic normality of the OLS estimator – and its large sample properties at large.

Large sample properties of OLS (1/2)

Theorem 1

Large Sample properties of the OLS Estimator. *Under Assumptions 1-6 the OLS estimator is consistent, that is:*

$$\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \boldsymbol{\beta}_0$$

and asymptotically normal, that is:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{K}_0^{-1}\boldsymbol{\Xi}_0\mathbf{K}_0^{-1}\right)$$

hence its asymptotic distribution is, for a given N, as follows.

$$\widehat{\boldsymbol{\beta}}_{OLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{1}{N}\mathbf{K}_0^{-1}\boldsymbol{\Xi}_0\mathbf{K}_0^{-1}\right)$$

Proof.

(Continues...)

Large sample properties of OLS (2/2)

Theorem 1

Proof.

(Continued.) Consistency follows from Assumptions 1-4 and the by fact that the residual term of the OLS decomposition vanishes at the probability limit. Asymptotic normality follows from "rephrasing" the earlier Central Limit Theorem result in terms of the random sequence:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}_0\right)=\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{x}_i\varepsilon_i$$

which, under Assumptions 1-6, gives – by Slutskij's Theorem and the Cramér-Wold Device – the stated limiting result. □

- This is worth to reiterate: like other asymptotic results, this one does not hinge upon particular distributional assumptions.
- It is the basis for the modern treatment of regression analysis in econometrics.

Robust variance-covariance estimation

Theorem 1 motivates the so-called **"robust"** *estimator* of the OLS asymptotic variance-covariance. It is **consistent**.

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) =$$

$$= \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right]^{-1} \left[\sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{OLS}\right)^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right] \left[\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right]^{-1}$$

- This estimator is "robust to heteroscedasticity." Also called **Huber-Eicker-White** or **heteroscedasticity-consistent** variance-covariance estimator of OLS, is the typical default choice in applied econometric practice.

- The estimator obtains through the Law of Large Numbers and the Continuous Mapping Theorem. Notice that the ε_i^2 terms are substituted with their squared residuals e_i^2: this works asymptotically (Eicker, 1967).

The homoscedastic case

In the special case of homoscedasticity, $\boldsymbol{\Xi}_0 = \sigma_0^2 \mathbf{K}_0$ and hence the asymptotic variance-covariance becomes:

$$\widehat{\boldsymbol{\beta}}_{OLS} \overset{A}{\sim} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{\sigma_0^2}{N}\mathbf{K}_0^{-1}\right)$$

and it is consistently estimated as follows.

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)^2 \left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}$$

- This is closer to the classical formula, still adopted in small sample analysis.

- Notice: the larger the variance-covariance of the regressors, the smaller the estimated variance-covariance of OLS. This follows since there is a larger support to fit the hyperplane.

Testing single hypotheses

These results enable statistical inference on the OLS estimator. Suppose that interest falls on a simple hypothesis like:

$$\mathrm{H}_0: \ \beta_{k0} = \mathrm{c}_k \qquad\qquad \mathrm{H}_1: \ \beta_{k0} \neq \mathrm{c}_k$$

where β_{k0} is an element of $\boldsymbol{\beta}_0$ and c_k is often zero. Then:

$$\mathrm{t}_{\mathrm{H}_0} = \frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\beta}_{k,OLS}\right)}} \xrightarrow{d} \mathcal{N}(0,1)$$

is an appropriate t-statistic for performing this test.

- The quantity in the denominator, is called **standard error** of $\widehat{\beta}_{k,OLS}$: it is the square root of the k-th diagonal element of the estimated variance-covariance of OLS.

Testing composite linear hypotheses

Suppose one is interested in L multiple **linear** hypotheses:

$$\mathrm{H}_0 : \ \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{c} \qquad\qquad \mathrm{H}_1 : \ \mathbf{R}\boldsymbol{\beta}_0 \neq \mathbf{c}$$

where $\mathbf{R}$ is a $L \times K$ matrix of full row rank L, while $\mathbf{c}$ is a $L \times 1$ vector. Note that, by the properties of the multivariate normal:

$$\sqrt{N}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{R}\left[\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)\right]\mathbf{R}^{\mathrm{T}}\right)$$

and hence these L hypotheses are *simultaneously* tested through the so-called **Wald statistic**.

$$\mathrm{W}_{\mathrm{H}_0} = \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)^{\mathrm{T}}\left[\mathbf{R}\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)\mathbf{R}^{\mathrm{T}}\right]^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)$$

As a particular case of a Hotelling's t-squared statistic, the Wald statistic asymptotically converges to a chi-square distribution.

$$\mathrm{W}_{\mathrm{H}_0} \xrightarrow{d} \chi^2_L$$

Small sample properties: a summary

- In small samples, asymptotic properties cannot be applied.
- One shall thus rely on **exact** distributional results in order to conduct statistical inference.
- Specifically, **unbiasedness** holds under Assumptions 1-6.
- However, deriving the variance-covariance of OLS requires **additional assumptions**.
- Furthermore, it is obtained **conditionally** on $\mathbf{X}$.
- Departures from these assumptions call for an extension of OLS: the **Generalized Least Squares** (GLS) estimator.

Unbiasedness in small samples

Small sample properties are best derived using compact matrix notation. Unbiasedness is obtained as follows.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{OLS}\right] &= \boldsymbol{\beta}_0 + \mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X}\right]\right] \\
&= \boldsymbol{\beta}_0 + \mathbb{E}_{\boldsymbol{X}}\left[\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\underbrace{\mathbb{E}\left[\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X}\right]}_{=\mathbf{0}}\right] \\
&= \boldsymbol{\beta}_0
\end{aligned}$$

Note the application of the Law of Iterated Expectation. The key assumptions are:

- 1 (linearity): it enables the OLS decomposition;
- 4 (exogeneity): it sets the inner expectation at zero.

Conditional covariance in small samples

Under the running assumptions, one can only derive the small sample variance-covariance of the OLS estimator *conditionally on the observed regressors*. Specifically:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right] &= \mathbb{E}\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\middle|\mathbf{X}\right] \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\end{aligned}$$

The key assumptions are:

- 1 (linearity): it enables the OLS decomposition;
- 4 (heteroscedasticity): it provides an expression for $\boldsymbol{\Sigma}$.

One could obtain the *unconditional* variance-covariance of OLS by averaging the above matrix over $\boldsymbol{X}$, in principle.

Spherical normal errors

Since $\boldsymbol{\Sigma}$ is usually unknown, the above expression is *unfeasible*. Thus, more assumptions are required.

Assumption 7

Spherical Errors. The error terms are homoscedastic, that is: $\sigma^2(\boldsymbol{x}_i) = \mathbb{V}\mathrm{ar}[\varepsilon_i | \boldsymbol{x}_i] = \sigma_0^2$; equivalently, $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right] = \sigma_0^2 \mathbf{I}$.

Assumption 8

Conditionally Normal Errors. The error term follows, given a regressor matrix $\mathbf{X}$, a conditionally normal distribution.

Together, these assumptions can be expressed as follows.

$$\boldsymbol{\varepsilon} | \mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma_0^2 \mathbf{I}\right)$$

Observe that the regressors are stochastic, and all these expressions are *conditional* on their realizations.

Gauss-Markov Theorem (1/2)

Under Assumptions 1-7, the conditional variance-covariance is:

$$\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\,\mathbf{X}\right]=\sigma_0^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

an expression associated with a celebrated result.

Theorem 2

Gauss-Markov Theorem. *Consider the linear regression model under Assumptions 1-7. Within the class of all linear, unbiased estimators defined as:*

$$\mathbb{B}=\left\{\widetilde{\boldsymbol{\beta}}=\mathbf{B}_0\mathbf{y}:\quad\mathbb{E}\left[\mathbf{B}_0\boldsymbol{y}\middle|\,\mathbf{X}\right]=\mathbf{B}_0\mathbf{X}\boldsymbol{\beta}_0+\mathbf{B}_0\,\mathbb{E}\left[\boldsymbol{\varepsilon}\middle|\,\mathbf{X}\right]=\boldsymbol{\beta}_0\right\}$$

the OLS estimator is the element of $\mathbb{B}$ *that delivers the minimum variance estimate of any element of* $\boldsymbol{\beta}_0$*, as well as of all possible linear combinations* $\mathbf{l}^{\mathrm{T}}\boldsymbol{\beta}_0$ *of* $\boldsymbol{\beta}_0$*, where* $\mathbf{l}$ *is a* $K\times 1$ *vector.*

Proof.

(Continues...)

Gauss-Markov Theorem (2/2)

Theorem 2

Proof.

By the definition of $\mathbb{B}$ and by Assumption 4, it holds that $\mathbf{B}_0\mathbf{X} = \mathbf{I}$ for all estimators in $\mathbb{B}$. Define $\mathbf{B}_1 \equiv \mathbf{B}_0 - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$ and observe that:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widetilde{\boldsymbol{\beta}}\middle|\mathbf{X}\right] &= \mathbf{B}_0\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\mathbf{X}\right]\mathbf{B}_0^{\mathrm{T}} \\
&= \sigma^2\left[\mathbf{B}_1 + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]\left[\mathbf{B}_1 + \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right]^{\mathrm{T}} \\
&= \sigma^2\left(\mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right) + \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{B}_1\mathbf{B}_1^{\mathrm{T}}\right) + \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}
\end{aligned}$$

where the third line follows from $\mathbf{B}_1\mathbf{X} = \mathbf{B}_0\mathbf{X} - \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{0}$. Thus:

$$\mathbf{l}^{\mathrm{T}}\,\mathbb{V}\mathrm{ar}\left[\widetilde{\boldsymbol{\beta}}\middle|\mathbf{X}\right]\mathbf{l} \geq \mathbf{l}^{\mathrm{T}}\,\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right]\mathbf{l}$$

which proves the result conditionally on $\mathbf{X}$; the unconditional result is obtained by taking the expectation over the random matrix $\boldsymbol{X}$. □

Towards small sample inference

- The Gauss Markov Theorem earns OLS the denomination **BLUE**: **Best Linear Unbiased Estimator**.

- This result is less relevant under heteroscedasticity, but it is still useful as a benchmark.

- In order to conduct statistical inference on OLS, note that under Assumption 8 it is:

$$\widehat{\boldsymbol{\beta}}_{OLS} \Big| \mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \sigma_0^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$$

  however, this expression requires knowledge of σ_0^2.

- One would be tempted to estimate the variance-covariance:

$$\widehat{\widehat{\mathbb{V}\mathrm{ar}}}\left[\widehat{\boldsymbol{\beta}}_{OLS} \Big| \mathbf{X}\right] = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

  while consistent, this estimator is however **biased**.

Expectation of the sum of squared residuals

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}\middle|\mathbf{X}\right] &= \mathbb{E}\left[\boldsymbol{y}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{y}\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}_0\right)^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\left(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}_0\right)\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\right)\middle|\mathbf{X}\right] \\
&= \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right)\middle|\mathbf{X}\right] \\
&= \mathrm{Tr}\left(\mathbf{M}_{\mathbf{X}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\mathbf{X}\right]\right) \\
&= \mathrm{Tr}\left(\sigma_0^2\mathbf{M}_{\mathbf{X}}\right) \\
&= \sigma_0^2\mathrm{Tr}\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right) \\
&= \sigma_0^2\mathrm{Tr}\left(\mathbf{I}\right)-\sigma_0^2\mathrm{Tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right) \\
&= \sigma_0^2\left(N-K\right)
\end{aligned}
$$

Degrees of freedom correction

The above derivation shows that the **unbiased** estimator of the conditional variance-covariance is:

$$\widehat{\mathbb{V}\text{ar}}\left[\widehat{\boldsymbol{\beta}}_{OLS}\middle|\mathbf{X}\right] = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{N-K}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

which is similar to the large sample version for homoscedasticity except that it applies a "degrees of freedom correction" with the multiplicative factor $N/(N-K)$.

The analysis of the sum of the squared residuals also allows to establish, under Assumption 8, that:

$$\left.\frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{\sigma_0^2}\right|\mathbf{X} = \left.\frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}}{\sigma_0^2}\right|\mathbf{X} \sim \chi^2_{N-K}$$

where the chi-square distribution here has a number of degrees of freedom: $N-K$, equal to the rank of matrix $\mathbf{M}_{\mathbf{X}}$.

Testing single hypotheses exactly

Tests of hypotheses are slightly more elaborate in small samples. Let again $\mathrm{H}_0: \ \beta_{k0} = \mathrm{c}_k$ and $\mathrm{H}_1: \ \beta_{k0} \neq \mathrm{c}_k$; the statistic

$$\mathrm{t}^*_{\mathrm{H}_0} \big| \, \mathbf{X} = \left| \frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\sigma_0^2 \, \widetilde{\mathrm{x}}_{kk}}} \right| \mathbf{X} \sim \mathcal{N}(0,1)$$

where $\widetilde{\mathrm{x}}_{kk}$ is the k-th diagonal element of the inverse of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, is again *unfeasible* without knowledge of σ_0^2. Let, however:

$$\mathrm{t}_{\mathrm{H}_0} = \sqrt{\sigma_0^2}\sqrt{N-K} \, \frac{\mathrm{t}^*_{\mathrm{H}_0}}{\sqrt{\mathbf{e}^{\mathrm{T}}\mathbf{e}}} = \sqrt{N-K} \, \frac{\widehat{\beta}_{k,OLS} - \mathrm{c}_k}{\sqrt{\mathbf{e}^{\mathrm{T}}\mathbf{e} \cdot \widetilde{\mathrm{x}}_{kk}}}$$

and since $\mathrm{t}^*_{\mathrm{H}_0}$ and $\mathbf{e}^{\mathrm{T}}\mathbf{e}/\sigma_0^2$ are independent conditionally on $\mathbf{X}$, the following conditional distributional result holds.

$$\mathrm{t}_{\mathrm{H}_0} \big| \, \mathbf{X} \sim \mathcal{T}_{N-K}$$

Testing composite linear hypotheses exactly

For composite linear hypotheses, the Wald statistic

$$\mathrm{W}^{*}_{\mathrm{H}_0} = \left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)^{\mathrm{T}} \frac{\left[\mathbf{R}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{R}^{\mathrm{T}}\right]^{-1}}{\sigma_0^2}\left(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{c}\right)$$

is such that $\mathrm{W}^{*}_{\mathrm{H}_0}\Big|\,\mathbf{X} \sim \chi^2_L$, but it is again unfeasible. One should instead use an F-statistic:

$$\mathrm{F}_{\mathrm{H}_0} = \mathrm{W}^{*}_{\mathrm{H}_0}\left(\frac{L}{N-K}\frac{\mathbf{e}^{\mathrm{T}}\mathbf{e}}{\sigma_0^2}\right)^{-1}$$

hence, by Observation 3 in Lecture 3 the following conditional distributional result holds.

$$\mathrm{F}_{\mathrm{H}_0}|\,\mathbf{X} \sim \mathcal{F}_{L,N-K}$$

This statistic is the basis for the so-called **F model test** about significance of the entire regression (with $\mathbf{R} = \mathbf{I}$ and $\mathbf{c} = \mathbf{0}$).

Generalizing the linear model

- The small sample properties are not fully satisfactory as they are restricted to homoscedasticity.

- What if the data are heteroscedastic, but $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}} \middle| \boldsymbol{X}\right]$ is known? Let

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\boldsymbol{\beta}_0 + \widetilde{\boldsymbol{\varepsilon}}$$

be a "generalized" linear regression model, where

$$\widetilde{\mathbf{y}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{y}; \quad \widetilde{\mathbf{X}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}; \quad \widetilde{\boldsymbol{\varepsilon}} \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}$$

and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is the following matrix such that $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} = \boldsymbol{\Sigma}$.

$$\boldsymbol{\Sigma}^{-\frac{1}{2}} \equiv \begin{bmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^{-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{\sigma_1^2}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\sigma_2^2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\sigma_N^2}} \end{bmatrix}$$

Generalized Least Squares

- The OLS estimator in this transformed model is called the **Generalized Least Squares** (GLS) estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{GLS} &= \left(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{y}} \\ &= \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ &= \boldsymbol{\beta}_0 + \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}\end{aligned}$$

- This estimator is **homoscedastic by construction**.

$$\mathbb{E}\left[\widetilde{\boldsymbol{\varepsilon}}\widetilde{\boldsymbol{\varepsilon}}^{\mathrm{T}}\middle|\boldsymbol{X}\right] = \boldsymbol{\Sigma}^{-\frac{1}{2}}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\middle|\boldsymbol{X}\right]\boldsymbol{\Sigma}^{-\frac{1}{2}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$$

- Its conditional variance-covariance, which is as follows, is efficient – by an extension of the Gauss-Markov theorem.

$$\widehat{\mathbb{V}\mathrm{ar}}\left[\widehat{\boldsymbol{\beta}}_{GLS}\middle|\mathbf{X}\right] = \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}$$

Feasibility of Generalized Least Squares

- Under Assumptions 1-8, the *exact* distribution of the GLS estimator is as follows – which enables statistical inference.

$$\widehat{\boldsymbol{\beta}}_{GLS} \Big| \mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}_0, \left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\right)$$

- The GLS estimator would be a solution to the limitations of small sample properties, if only matrix $\boldsymbol{\Sigma}$ were known.

- When $\boldsymbol{\Sigma}$ is unknown, GLS is said to be **unfeasible**.

- A **Feasible Generalized Least Squares** (FGLS) version of the estimator can be constructed through an **estimator** of $\boldsymbol{\Sigma}$. This requires several steps to be illustrated next.

- FGLS is based on assumptions about the **functional form** of the error term's variance conditional on the regressors.

Feasible Generalized Least Squares: procedure

1. Assume a **functional form** for $\sigma^2(\boldsymbol{x}_i) = \mathbb{V}\text{ar}\left[\varepsilon_i^2 \middle| \boldsymbol{x}_i\right]$, the dependence of the error term's variance on the regressors.

2. Estimate the **main regression** model of interest via OLS, which returns an **unbiased** and consistent estimate of $\boldsymbol{\beta}_0$, and calculate the *squared residuals* $(e_1^2, e_2^2, \ldots, e_N^2)$.

3. Estimate via OLS the **assumed** conditional variance model $\sigma^2(\boldsymbol{x}_i)$: the details depend on the assumed functional form.

4. Construct matrix $\widehat{\boldsymbol{\Sigma}}$, the estimate of $\boldsymbol{\Sigma}$, accordingly.

5. Finally, calculate the FGLS estimator as follows.

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

Feasible Generalized Least Squares: example

- Let the conditional variance be $\sigma^2(\boldsymbol{x}_i) = \exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\psi}\right)$.

- An estimator for $\boldsymbol{\psi}$ is obtained at point 3 of the procedure by running the following model on the point 2 residuals.

$$\log e_i^2 = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\psi} + \varpi_i$$

  Here ϖ_i is an error term with $\mathbb{E}[\varpi_i | \boldsymbol{x}_i] = 0$.

- Matrix $\widehat{\boldsymbol{\Sigma}}$ is thus obtained as follows.

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \exp\left(\mathbf{x}_1^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) & 0 & \dots & 0 \\ 0 & \exp\left(\mathbf{x}_2^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \exp\left(\mathbf{x}_N^{\mathrm{T}}\widehat{\boldsymbol{\psi}}_{OLS}\right) \end{bmatrix}$$

Weighted Least Squares

- By denoting as $\widehat{\sigma}_i$ the square root of the i-th element of the diagonal of $\widehat{\boldsymbol{\Sigma}}$, FGLS can be equivalently obtained via OLS run on the following transformed model.

$$\frac{y_i}{\widehat{\sigma}_i} = \beta_1 \frac{x_{i1}}{\widehat{\sigma}_i} + \beta_2 \frac{x_{i2}}{\widehat{\sigma}_i} + \cdots + \beta_K \frac{x_{iK}}{\widehat{\sigma}_i} + \frac{\varepsilon_i}{\widehat{\sigma}_i}$$

  Possibly, it is $x_{i1} = 1$ for all $i = 1, \ldots, N$.

- This approach is called **Weighted Least Squares** (WLS).

- If $\sigma^2(\boldsymbol{x}_i)$ is correctly specified, the FGLS-WLS estimator is the most efficient one under heteroscedasticity, even better than "robust" OLS in large samples.

- But if the $\sigma^2(\boldsymbol{x}_i)$ model is wrong, FGLS-WLS might be *less* efficient than standard OLS. Thus, in large samples robust OLS is the preferred conservative option.

Introducing dependent observations

- Both large and small sample approaches discussed thus far assume **independent observations** (Assumption 2).
- That is, $\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{x}_i, \boldsymbol{x}_j\right] = 0$ for any two observations i & j.
- In large samples, this gives a convenient expression for $\boldsymbol{\Xi}_0$.
- This assumption is, however, unlikely to hold in many real world scenarios of interest.
- This jeopardizes statistical inference performed on $\widehat{\boldsymbol{\beta}}_{OLS}$!
- A taxonomy of typical cases of **dependent observations** is summarized next.
- Then, typical solutions for this problem are introduced.

Autocorrelation in time

- Traditionally, econometrics was very focused on time series.

- Time series regression models would be expressed as:

$$y_t = \mathbf{x}_t^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_t$$

  here the subscript t denotes time.

- The phrase **autocorrelation in time** expresses the notion that the current error term ε_t statistically depends to those from the past and into the future:

$$\mathbb{E}\left[\varepsilon_t \varepsilon_{t-s} \middle| \boldsymbol{x}_t, \boldsymbol{x}_{t-s}\right] \neq 0$$

  where $s \neq 0$. This is a typical feature of time series.

Spatial correlation

- Correlation can extend in space as well as in time.

- Suppose that in a cross-sectional setting, the observations indexed as $i = 1, \ldots, N$ are reciprocally related through a pairwise **distance measure** $d_{ij} \geq 0$.

- One can give multiple interpretations to d_{ij}: like physical or geographical distance, network distance, *et cetera*.

- Under **spatial correlation**, cross-observation dependence is a function of their pairwise distance.

$$\mathbb{E}\left[\varepsilon_i \varepsilon_j \middle| \boldsymbol{x}_i, \boldsymbol{x}_j\right] = g\left(d_{ij}\right) \neq 0$$

- Intuitively, observations "close" to one another are affected by similar circumstances.

Within-group correlation (1/2)

- Let the sample be split by a number $C < N$ of **groups** or **clusters** indexed by $c = 1, \ldots, C$; each observation belongs to *one and only one* group or cluster of size N_c.

- Observations are easily indexed by their group or cluster.

$$y_{ic} = \mathbf{x}_{ic}^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_{ic}$$

- Under **within-group correlation** error terms are split as follows.

$$\varepsilon_{ic} = \alpha_c + \epsilon_{ic}$$

- While α_c is the **group** or **cluster shock**, which captures the shared component of the error term...

- ... ϵ_{ic} is the observation-specific **idiosyncratic shock**.

Within-group correlation (2/2)

- Formally, ϵ_{ic} is independent across pairs of observations:

$$\mathbb{E}\left[\epsilon_{ic}\epsilon_{jg}\middle|\,\boldsymbol{x}_{ic},\boldsymbol{x}_{jg}\right]\begin{cases}=\sigma_{\epsilon}^{2}\left(\boldsymbol{x}_{i}\right) & \text{if } i=j\\ =0 & \text{if } i\neq j\end{cases}$$

- . . . while α_c correlates within groups, but not across groups.

$$\mathbb{E}\left[\alpha_{c}\alpha_{g}\middle|\,\boldsymbol{x}_{ic},\boldsymbol{x}_{jg}\right]\begin{cases}\neq 0 & \text{if } c=g\\ =0 & \text{if } c\neq g\end{cases}$$

- This setup is suited to describe similar "shocks" that affect groups of individuals (classmates, compatriots) or different "categories" (firms in the same industry, cities in the same administrative unit and so on).

Combinations of the above

- All these scenarios can co-exist at the same time.

- Suppose you have a panel where $i = 1, \ldots, N$ indexes panel units, $t = 1, \ldots, T$ indexes time, while $c = 1, \ldots, C$ indexes groups or clusters.

- The regression model would write as follows.

$$y_{itc} = \mathbf{x}_{itc}^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_{itc}$$

- Autocorrelation in time, spatial correlation and in addition within-group correlation can all be present.

- As a special case, panel units and groups coincide ($C = N$): the group shocks α_c would then be called **random effects** and represent factors that are constant to a panel unit.

Solutions to dependent observations

- Several solutions to those issues are available.
- In small and large samples alike, it is possible to adapt the GLS framework for this purpose. However, this relies upon making the right assumptions on $\boldsymbol{\Sigma}$ for efficiency's sake.
- In large samples, one must account for the fact that matrix $\boldsymbol{\Xi}_0$ has a more general form, which can be estimated.
- A common approach to address within-group correlation is **clustered covariance estimation** (CCE).
- More generally, the **heteroscedasticity-autocorrelation consistent** (HAC) approach can address autocorrelation in time as well as spatial correlation.
- Both CCE and HAC deliver asymptotic estimators for $\boldsymbol{\Xi}_0$.

A GLS solution for autocorrelation in time

- All GLS approaches to dependent observations rely on the fact that $\mathbf{\Sigma}$ is semidefinite positive and can factorized as:

$$\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Sigma}^{-\frac{1}{2}} = \mathbf{I}$$

  even if it is not diagonal, resulting in a suitable matrix $\mathbf{\Sigma}^{\frac{1}{2}}$.

- Let ε_t be autocorrelated in time according to a **first-order autoregressive** – AR(1) – process:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \xi_t$$

  where $|\rho| < 1$ and ξ_t is an i.i.d. shock with variance σ^2.

- Here FGLS is about estimating ρ, since $\mathbf{\Sigma}$ is as follows.

$$\mathbf{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho^T \\ \rho & 1 & \dots & \rho^{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^T & \rho^{T-1} & \dots & 1 \end{bmatrix}$$

A GLS solution for clustered correlation (1/2)

- Consider *constant* correlation within groups:

$$\mathbb{E}\left[\alpha_c \alpha_g \middle| \boldsymbol{x}_{ic}, \boldsymbol{x}_{jg}\right] \begin{cases} = \sigma_\alpha^2 & \text{if } c = g \\ = 0 & \text{if } c \neq g \end{cases}$$

  a model called **cluster-specific random effects** (CSRE).

- If in addition the idiosyncratic shock is homoscedastic, i.e.

$$\sigma_\epsilon^2\left(\boldsymbol{x}_i\right) = \sigma_\epsilon^2$$

  for all $i = 1, \ldots, N$, then $\boldsymbol{\Sigma}$ is **block-diagonal**.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_C \end{bmatrix}$$

A GLS solution for clustered correlation (2/2)

- Specifically, for $c = 1, \ldots, C$ it is:

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \sigma_\epsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & \ldots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\epsilon^2 + \sigma_\alpha^2 & \ldots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \ldots & \sigma_\epsilon^2 + \sigma_\alpha^2 \end{bmatrix} = \sigma_\epsilon^2 \mathbf{I} + \sigma_\alpha^2 \iota \iota^{\mathrm{T}}$$

  where $\mathbf{I}$ is a suitable identity matrix and ι is a unit vector of the same dimension as the size N_c of some cluster c.

- Here, FGLS amounts to estimating the variance-covariance parameters σ_α^2 and σ_ϵ^2 with the preliminary OLS estimates.

- This is the approach typically followed to address **random effects** in panel data models.

- This may be inefficient if $\boldsymbol{\Sigma}$ is not of the assumed structure.

Clustered Covariance Estimation (1/5)

- To introduce cluster covariance estimation, some additional "semi-compact" notation is necessary.

- Consider a single group or **cluster** c, index its observations as $i = 1, \ldots, N_c$, and stack them vertically as follows.

$$\mathbf{y}_c = \begin{bmatrix} y_{1c} \\ y_{2c} \\ \vdots \\ y_{N_c c} \end{bmatrix}; \quad \mathbf{X}_c = \begin{bmatrix} \mathbf{x}_{1c}^{\mathrm{T}} \\ \mathbf{x}_{2c}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{N_c c}^{\mathrm{T}} \end{bmatrix}; \quad \boldsymbol{\varepsilon}_c = \begin{bmatrix} \varepsilon_{1c} \\ \varepsilon_{2c} \\ \vdots \\ \varepsilon_{N_c c} \end{bmatrix}$$

where, more specifically, $\mathbf{X}_c$ is as follows.

$$\mathbf{X}_c = \begin{bmatrix} \mathbf{x}_{1c}^{\mathrm{T}} \\ \mathbf{x}_{2c}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{N_c c}^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} x_{11c} & x_{21c} & \cdots & x_{K1c} \\ x_{12c} & x_{22c} & \cdots & x_{K2c} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1N_c c} & x_{2N_c c} & \cdots & x_{KN_c c} \end{bmatrix}$$

Clustered Covariance Estimation (2/5)

- In this environment, the linear model can also be expressed as follows:

$$\mathbf{y}_c = \mathbf{X}_c \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_c$$

holding over all clusters $c = 1, \ldots, C$.

- Note that these C groups are allowed to have different sizes N_c; thus, the OLS estimator can be expressed in one of the following three equivalent ways.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{OLS} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\
&= \left(\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c\right)^{-1}\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{y}_c \\
&= \left(\sum_{c=1}^{C}\sum_{i=1}^{N_c}\mathbf{x}_{ic}\mathbf{x}_{ic}^{\mathrm{T}}\right)^{-1}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\mathbf{x}_{ic}y_{ic}
\end{aligned}$$

Clustered Covariance Estimation (3/5)

- With group correlation, standard inference is invalid since:

$$\begin{aligned}
\boldsymbol{\Xi}_0 &= \lim_{N\to\infty} \mathbb{V}\mathrm{ar}\left[\frac{1}{\sqrt{N}}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\boldsymbol{x}_{ic}\varepsilon_{ic}\right] \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{c=1}^{C}\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{N_c}\boldsymbol{x}_{ic}\varepsilon_{ic}\right] \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}\mathbb{E}\left[\varepsilon_{ic}\boldsymbol{x}_{ic}\boldsymbol{x}_{jc}^{\mathrm{T}}\varepsilon_{jc}\right]
\end{aligned}$$

  but this expression can be reduced no further.

- Under appropriate assumptions, a Central Limit Theorem *for dependent observations* can still ensure the following.

$$\frac{1}{\sqrt{N}}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\mathbf{x}_{ic}\varepsilon_{ic} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Xi}_0\right)$$

Clustered Covariance Estimation (4/5)

- An appropriate, consistent estimator of $\boldsymbol{\Xi}_0$ is necessary:

$$\widehat{\boldsymbol{\Xi}}_{CCE} = \frac{1}{N}\sum_{c=1}^{C}\mathbf{X}_c^{\mathrm{T}}\mathbf{e}_c\mathbf{e}_c^{\mathrm{T}}\mathbf{X}_c =$$

$$= \frac{1}{N}\sum_{c=1}^{C}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}\left(y_{ic} - \mathbf{x}_{ic}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)\mathbf{x}_{ic}\mathbf{x}_{jc}^{\mathrm{T}}\left(y_{jc} - \mathbf{x}_{jc}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{OLS}\right)$$

  where, similarly to the standard case:

$$\mathbf{e}_c \equiv \mathbf{y}_c - \mathbf{X}_c\widehat{\boldsymbol{\beta}}_{OLS}$$

  replaces $\boldsymbol{\varepsilon}_c$ without invalidating the asymptotics.

- As hinted, this estimator is consistent: $\widehat{\boldsymbol{\Xi}}_{CCE} \xrightarrow{p} \boldsymbol{\Xi}_0$.

- Intuitively, observations in the same cluster interact in the variance-covariance estimation.

Clustered Covariance Estimation (5/5)

The CCE or **cluster-robust** estimator of the OLS asymptotic variance-covariance is thus obtained as follows.

$$\widehat{\mathbb{A}\mathrm{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \left[\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}} \mathbf{X}_c\right]^{-1} \left[\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}} \mathbf{e}_c \mathbf{e}_c^{\mathrm{T}} \mathbf{X}_c\right] \left[\sum_{c=1}^{C} \mathbf{X}_c^{\mathrm{T}} \mathbf{X}_c\right]^{-1}$$

- With a few clusters C, a multiplicative "degree of freedom correction" $\frac{C}{C-1}\frac{N}{N-K}$ can lead to more precise inferences.
- If C is between 20 and 50, the current practice favors tests performed against the t- and F-distributions.
- Clustering is ubiquitous in current empirical research; with panel data, one should *at least* cluster by panel units.
- In proper experimental studies however this is unnecessary: by construction, $\boldsymbol{x}_{ic}$ is independent of ε_{ic}, thus $\boldsymbol{\Xi}_0 = \sigma_0^2 \mathbf{K}_0$.

Multi-way clustering

- Group correlation might occur along multiple dimensions. The CCE estimator must be thus adjusted accordingly.

- In panel data, say, one can think of two group dimensions: panel units and *time* (units receive simultaneous correlated shocks). A proper **two-way clustering** formula would be:

$$\widehat{\mathbb{A}\text{var}}_{\mathbb{I},\mathbb{T}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = \\ = \widehat{\mathbb{A}\text{var}}_{\mathbb{I}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) + \widehat{\mathbb{A}\text{var}}_{\mathbb{T}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) - \widehat{\mathbb{A}\text{var}}_{\mathbb{J}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)$$

  which combines the CCE formulae where groups are based on panel units ($\mathbb{I}$), time ($\mathbb{T}$) or the interaction thereof, that is the unique unit-time observations ($\mathbb{J}$) respectively.

- Extensions to higher dimensions (**multi-way clustering**) also exist (Cameron et al., 2011).

Heteroscedasticity-autocorrelation consistency

- Like within-group dependence, autocorrelation in time and spatial correlation both imply more general versions of $\boldsymbol{\Xi}_0$.

- All HAC estimators are based upon $K \times K$ matrices $\widehat{\boldsymbol{\Xi}}_{HAC}$ that consistently estimate $\boldsymbol{\Xi}_0$.

$$\widehat{\boldsymbol{\Xi}}_{HAC} \xrightarrow{p} \boldsymbol{\Xi}_0$$

- Similarly to CCE, the OLS asymptotic variance-covariance is then estimated as:

$$\widehat{\mathbb{A}\text{var}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right) = N\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}\widehat{\boldsymbol{\Xi}}_{HAC}\left[\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right]^{-1}$$

  where N (or possibly T) usually simplifies with N^{-1} (T^{-1}) found in the expression of $\widehat{\boldsymbol{\Xi}}_{HAC}$.

- HAC estimators are more flexible but less transparent than CCE (which can be at times seen as a special case of HAC).

The Newey-West HAC estimator (1/2)

- The original HAC estimator was introduced by Newey and West (1987) for autocorrelated time series in linear models like $y_t = \mathbf{x}_t^{\mathrm{T}} \boldsymbol{\beta}_0 + \varepsilon_t$ (with $t = 1, \ldots, T$). It is based on:

$$\widehat{\boldsymbol{\Xi}}_{NW} = \sum_{s=-(T-1)}^{T-1} \kappa_T(s) \frac{1}{T} \sum_{t=1}^{T} e_t \mathbf{x}_t \mathbf{x}_{t+s}^{\mathrm{T}} e_{t+s}$$

  where NW meaning "Newey-West," $e_t = y_t - \mathbf{x}_t^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_{OLS}$ and e_{t+s} is similarly defined.

- Here $\kappa_T(s)$ is a **weighting kernel** that is decreasing in $|s|$, and such that $\kappa_T(s) = 0$ if $t + s < 1$ or $t + s > T$.

- Similarly to CCE, correlated observations interact in $\widehat{\boldsymbol{\Xi}}_{NW}$. For it to be consistent and a proper Central Limit Theorem to apply, the kernel must "decay" sufficiently rapidly.

The Newey-West HAC estimator (2/2)

- The most popular weighting kernel is the **Bartlett kernel**:

$$\kappa_{B_T}(s) = \left(1 - \frac{|s|}{B_T}\right)^+$$

  where $2B_T$ is the **base** of the kernel; notice that $\kappa_T(0) = 1$ and that it decreases uniformly for higher values of $|s|$ until $\kappa_T(|\widetilde{s}|) = 0$ for $\widetilde{s} \geq B_T$.

- There is some empirical tension: $2B_T$ must be long enough to capture the relevant autocorrelation, but not too long.

- In a panel model like $y_{it} = \mathbf{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_{it}$, $\widehat{\boldsymbol{\Xi}}_{NW}$ extends as:

$$\widehat{\boldsymbol{\Xi}}_{NW} = \sum_{s=-(T-1)}^{T-1} \kappa_T(s) \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} e_{it}\mathbf{x}_{it}\mathbf{x}_{i(t+s)}^{\mathrm{T}} e_{i(t+s)}$$

  where $\kappa_T(s)$ can extend over the entire panel length (notice that if $\kappa_T(s) = 1$ always, this coincides with panel CCE).

Spatial correlation consistency

- This estimator can extend to spatial correlation. A version that is heteroscedasticity and **spatial** correlation consistent (HSC) is, given pairwise observation distances d_{it}:

$$\widehat{\mathbf{\Xi}}_{HSC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa_N\left(d_{ij}\right)e_i\mathbf{x}_i\mathbf{x}_j^{\mathrm{T}}e_j$$

  where the kernel $\kappa_N\left(d_{ij}\right)$ must decay sufficiently rapidly in d_{ij}. This estimator was first studied by Conley (1999).

- In panel data where autocorrelation in time and space can co-exist, an alternative to two-way clustering is a "HASC" estimator, which proceeds as follows.

$$\widehat{\mathbf{\Xi}}_{HASC} =$$

$$= \sum_{s=-(T-1)}^{T-1}\kappa_T\left(s\right)\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{N}\kappa_N\left(d_{ij}\right)e_{it}\mathbf{x}_{it}\mathbf{x}_{j(t+s)}^{\mathrm{T}}e_{j(t+s)}$$