Assumptions of the **classical linear regression model**, what to do when they are violated, and **estimator properties**

Contents

*Disclaimer: sections and lines in brown correspond to content which is **very much** 'under construction'.*

1 Assumptions of the CLRM for predictive inference

The classical linear regression model (CLRM) consists of a set of assumptions that describe how the dataset is produced by a data generating process (DGP). By decreasing order of importance[1]:

Notation:	System of n equations	Matrix
Model:	$y_i = \mathrm{x}_i'\beta + e_i \quad (i = 1, ..., n)$	$y = \mathbf{X}\beta + \mathrm{e}$
Assumptions		
A1. linearity	The model is linear in β	The model is linear in β
A2. identification	$\rho_{X_k, X_l} \approx 1$	$\mathbf{X}_{N\times K}$ has rank K
A3. exogeneity	$\mathbb{E}[e_i\|\mathbf{X}] = 0$	$\mathbb{E}[\mathrm{e}\|\mathbf{X}] = 0_{N\times 1}$
A4. spherical errors	$e_i\|\mathbf{X} \overset{\text{iid}}{\sim} (0, \sigma^2)$	$\mathbb{V}[\mathrm{e}\|\mathbf{X}] = \sigma^2\mathbf{I}_N$
– **independent errors**	$\mathrm{cov}[e_i, e_j\|\mathbf{X}] = 0$	
– **homoskedastic errors**	$\mathbb{V}[e_i\|\mathbf{X}] = \sigma^2$ ~~σ_i^2~~	
A5. normal errors	$e_i\|\mathbf{X} \sim \mathcal{N}(0, \sigma^2)$	$\mathrm{e}\|\mathbf{X} \sim \mathcal{N}(0_{N\times 1}, \sigma^2\mathbf{I}_N)$

(A1) **Linearity in the parameters** and correct model specification (notably an additive error term). I.e., the linear functional form coïncides with the actual DGP.

(A2) **Identification:** regressors are linearly independent (no perfect colinearity).
If this is violated, drop one X, or transform them into one X.

(A3) **Strict exogeneity of regressors**: all other factors that affect y are unrelated to X.
$\mathbb{E}[\mathrm{e}|\mathbf{X}] = 0$ also implies $\mathbb{E}[\mathrm{e}]$=0 and $\mathbb{E}[X'\mathrm{e}]$=0, leading to $\mathrm{cov}(e_i, X)$=0: X and e are uncorrelated.

(A4) **Spherical errors**

- **Independent errors** $\implies$ no autocorrelation: $\mathrm{cov}(e_i, e_j|X) = \mathbb{E}[e_i e_j|X] = 0$
  I.e., errors are randomly spread around the regression line.
  If this is violated, e.g., by serial correlation (likely with time series data), try taking lags of regressors, or switch to an autoregressive or a moving average model...
- **Homoskedastic errors**: equal *conditional* variance $\mathbb{V}[e_i|X] = \sigma^2$
  The error variance is a measure of model uncertainty. Homoskedasticity means uncertainty, i.e., the spread of errors, is identical across the support of y.
  If this is violated, $\hat{\beta}_{\text{OLS}}$ remains valid but isn't efficient – Weighted Least Squares has a lower variance. Look for omitted variables, remove outliers, perform a log-transformation...

(A5) **Normal errors**
This assumption is not required for estimating the regression but for making **inference**, e.g., computing confidence intervals or p-values. Without (A5), t and F tests are invalid.
If this is violated, we have to appeal to asymptotics: the properties of $\hat{\beta}$ for large samples. Indeed: the one-sample t-test for β, which tests the null hypothesis that β=0, assumes that the sampling distribution of $\hat{\beta}$ is normal. If errors are not normal, then $\hat{\beta}$ isn't normal. However, when n is large enough, Laws of Large Numbers (LLNs) and Central Limit Theorems (CLTs) say that the asymptotic sampling distribution of $\hat{\beta}$ is normal, s.t. t and F tests are robust to departures from normality if n is large.
If errors are highly non-normal (e.g., long tailed), appealing to an asymptotically normal approximation may be unreasonable, and one may want to consider an alternative (e.g., bootstrap).

[1] See Gelman et al. (2020, Chapter 11)

2 OLS and ML estimators of $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\sigma^2}\}$ and their statistical properties

There are often several possible estimators to estimate a relationship between X and y. How one chooses between them (besides their facility of computation) is motivated by their **statistical properties**.

2.1 Estimator properties

Let $\hat{\theta}$ be an estimator for the population parameter θ_0, based on a sample of size n. Note that $\hat{\theta}$ is a random variable itself. We can conceive various samples of size n, and thus a sequence of $\{\hat{\theta}\}$. The $\hat{\theta}$ estimator has:

- finite sample properties: how $\hat{\theta}$ behaves for a finite n. *Ex: bias, efficiency.*
- asymptotic properties: how $\hat{\theta}$ behaves as $n \to \infty$. *Ex: consistency, a known asymptotic distribution.*

Consistency and having an asymptotic distribution are desirable properties as they permit statistical inference at least in large samples. However, as we always deal with finite samples, the behaviour of estimators in finite samples may be more important. In effect, bias and efficiency are the most common selection criteria.

Finite sample properties

- $\hat{\theta}$ is **unbiased** iff $\mathbb{E}\big[\hat{\boldsymbol{\theta}}|\boldsymbol{X}\big] = \boldsymbol{\theta_0}$
  Taking *repeated samples* of size n, estimating the value of $\hat{\theta}$ for each, the average of these values equals the true (unknown) value of the parameter. I.e., the estimator is true *on average*. On the contrary, biased means $\hat{\theta}$ is systematically different from the true value (bias = systematic error).
  ⚠ *It does not mean that the estimate from any one sample is even close to the true parameter.*
- $\hat{\theta}$ is **efficient** iff it has the lowest possible variance of all estimators: $\mathbb{V}\big[\hat{\boldsymbol{\theta}}\big] \leqslant \mathbb{V}\big[\tilde{\boldsymbol{\theta}}_{...}\big]$
  The "best" estimator = the one with the smallest possible variance, i.e., that deviates as little as possible from the true value we are trying to estimate. (For an unbiased estimator, that variance corresponds to the Cramér-Rao Lower Bound.)

Asymptotic properties

- $\hat{\theta}$ is **asymptotically unbiased** iff $\mathbb{E}\big[\hat{\boldsymbol{\theta}}|\boldsymbol{X}\big] \xrightarrow[n\to+\infty]{p} \boldsymbol{\theta_0}$
- $\hat{\theta}$ is **asymptotically efficient** iff $\mathbb{V}\big[\hat{\boldsymbol{\theta}}|\boldsymbol{X}\big] \xrightarrow[n\to+\infty]{p}$ ***asymptotic Cramer Rao lower bound***
- $\hat{\theta}$ is **consistent** iff $\hat{\boldsymbol{\theta}}|\boldsymbol{X} \xrightarrow[n\to+\infty]{p} \boldsymbol{\theta_0}$
  Sufficient conditions are that $\hat{\theta}$ be asymptotically unbiased, and its variance shrink to 0 as $n \to \infty$.

Various estimators exist. In frequentist statistics, the Maximum Likelihood Estimator (MLE) and the Ordinary Least Squares (OLS) estimator are among the most common. They are described below.

A few preliminary remarks:
- OLS and ML are rooted in different mathematical disciplines: calculus for OLS, probabilities for ML. OLS makes no assumption on the probabilistic nature of the variables, it is deterministic.
- OLS is tailored to the linear regression model[2], and is often used because $\hat{\beta}_{\text{OLS}}$ is unbiased even with non-spherical errors. However, OLS may of course be inefficient, and we may have to fix the usual SEs.
- ML includes OLS as a special case: if $e|X \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 I_n)$, then $y|X \sim \text{MVN}(X\beta, \sigma^2 I_n)$ and $\hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{MLE}}$.

[2] Least squares is widely used because the estimated function $f(X_i, \hat{\beta})$ approximates the conditional expectation $\mathbb{E}[y|X]$.

2.2 ML estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}} = \left\{\hat{\boldsymbol{\beta}}_{\text{MLE}}, \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\text{MLE}}\right\}$

Definition The likelihood function in a regression model is the probability density of the data given the parameters and predictors. Assuming iid observations: $L\big(y \mid X, \theta\big) = f\big(X_1, ..., X_n, \theta\big) = f\big(X_1, \theta\big)...f\big(X_n, \theta\big) = \prod_{i=1}^{n} f\big(X_i, \theta\big)$. We can then also compute the loglikelihood: $\log L\big(y \mid X, \theta\big) = \sum_{i=1}^{n} \log f\big(X_i, \theta\big)$.

The Maximum Likelihood Estimator (MLE) is the value of θ s.t. under the assumed model, the observed data is most likely:

$$\hat{\theta}_{\text{MLE}} \equiv \underset{\theta}{\text{argmax}}\ L\big(y \mid X, \theta\big) = \underset{\theta}{\text{argmax}}\ \log L\big(y \mid X, \theta\big)$$

Solution (A5) $\implies y|X \sim \text{MVN}(X\beta, \sigma^2 I_n)$. We can write the likelihood or joint density of y: $f(y|X, \beta, \sigma^2) = (2\pi\sigma^2)^{\frac{-n}{2}} e^{-\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}}$, and thus the log-likelihood: $l \equiv \ln f(y|\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{(y-X\beta)'(y-X\beta)}{2\sigma^2}$. The two FOCs of the maximization problem give an exact closed-form solution[3]:

$$\begin{cases} \dfrac{\partial l}{\partial \beta} = 0 \iff \dfrac{-1}{2\hat{\sigma}^2}\big(-2X'y + 2X'X\hat{\beta}\big) = 0 \iff \hat{\beta}_{\text{MLE}} = (X'X)^{-1}X'y \\ \dfrac{\partial l}{\partial \sigma^2} = 0 \iff \dfrac{-n}{2\hat{\sigma}^2} + \dfrac{(y - X\hat{\beta})'(y - X\hat{\beta})}{2\hat{\sigma}^4} = 0 \iff n\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta}) \iff \hat{\sigma}^2_{\text{MLE}} = \dfrac{\hat{e}'\hat{e}}{n} = \dfrac{r'r}{n} \end{cases}$$

Properties [assuming (A1)–(A5)]

- Finite samples

  $\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$ is **unbiased** $\mathbb{E}\big[\hat{\beta}_{\text{MLE}}|X\big] = \mathbb{E}[(X'X)^{-1}X'y|X] = \mathbb{E}[(X'X)^{-1}X'(X\hat{\beta} + e)|X]$
  $= \mathbb{E}[\hat{\beta}|X] + \mathbb{E}[(X'X)^{-1}X'e|X] = \beta_0$

  efficient $\mathbb{V}\big[\hat{\beta}_{\text{MLE}}|X\big] = \mathbb{E}\left[\big(\hat{\beta} - \mathbb{E}\big[\hat{\beta}\big]\big)\big(\hat{\beta} - \mathbb{E}\big[\hat{\beta}\big]\big)' \mid X\right] = \mathbb{E}\left[\big(\hat{\beta} - \beta_0\big)\big(\hat{\beta} - \beta_0\big)' \mid X\right]$
  $= \mathbb{E}\big[(X'X)^{-1}X'e\big((X'X)^{-1}X'e\big)'|X\big]$
  $= (X'X)^{-1}X'\,\mathbb{E}[ee'|X]\,X(X'X)^{-1} = \sigma^2(X'X)^{-1} \leqslant \mathbb{V}\big[\hat{\beta}_{...}|X\big]$

  normally distributed $\hat{\beta}_{\text{MLE}} = \beta_0 + (X'X)^{-1}X'e \sim \mathcal{N}\left(\beta_0, \sigma^2(X'X)^{-1}\right)$

  $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MLE}}$ is *downward biased* $\mathbb{E}\big[\hat{\sigma}^2_{\text{MLE}}|X\big] = \frac{1}{n}\mathbb{E}[r_i'r_i|X] = ... = \frac{n-k}{n}\sigma^2 < \sigma^2$
  The variance is underestimated. The size of the bias will decrease as the sample size gets larger. To overcome this problem, we can compute the sample variance s^2 instead of $\hat{\sigma}^2_{\text{MLE}}$.

- Asymptotics

  $\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$ is **asymptotically unbiased** as is unbiased

  asymptotically efficient as is efficient

  consistent as 1. is asymptotically unbiased, and 2. $\mathbb{V}\big[\hat{\beta}_{\text{MLE}}|X\big] = ... \xrightarrow[n\to\infty]{p} 0$

  $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MLE}}$ is **asymptotically unbiased** as $\lim_{n\to\infty} \mathbb{E}\big[\hat{\sigma}^2_{\text{MLE}}|X\big] = \lim_{n\to\infty}\left(\sigma^2 - \frac{k}{n}\sigma^2\right) = \sigma^2$

  asymptotically efficient as $\sqrt{n}(\hat{\sigma}^2_{\text{MLE}} - \sigma^2) \xrightarrow{d} \mathcal{N}\big(0, 2\sigma^4\big)$

  consistent as 1. is asymptotically unbiased, and 2. $\mathbb{V}\big[\hat{\sigma}^2_{\text{MLE}}|X\big] = \frac{2\sigma^4(n-k)}{n^2} \xrightarrow[n\to\infty]{p} 0$

[3] The likelihood function must be differentiable in order to apply the derivative test for determining maxima. In some cases, the FOCs can be solved explicitly (e.g., the OLS estimator maximizes the likelihood of the linear regression model). Under most circumstances, however, numerical methods will be necessary to find the maximum of the likelihood function.

2.3 OLS estimator $\hat{\boldsymbol{\theta}}_{\text{OLS}} = \left\{\hat{\boldsymbol{\beta}}_{\text{OLS}}, \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\text{OLS}}\right\}$

Definition The fit of a model $y = g(X, \beta)$ to each data point is measured by its residual $r_i \equiv y_i - g(x_i, \beta)$. The Ordinary Least Squares (OLS) estimator computes, in the context of a model linear in the parameters $g(X, \beta) = \sum_{j=1}^{k} \beta_j h_j(X)$, the values of the parameters that minimize the sum of the squares of the residuals; it is the one that best fit the data.

$$\hat{\beta}_{\text{OLS}} \equiv \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} r_i^2$$

With the residuals $r_i \equiv \hat{e}_i$ from the fit, we compute as estimator of σ^2 the statistic $\hat{\sigma}^2_{\text{OLS}} \equiv s^2 \equiv \frac{r'r}{n-k} = \frac{\sum r_i^2}{n-k}$.

Solution The FOC of the minimization problem gives an exact closed-form solution (which the SOC guarantees is a minimum iff the matrix $X'X$ is positive definite):

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta_0 + e) = \beta_0 + (X'X)^{-1}X'e$$

Properties [assuming (A1)–(A3)]

- Finite samples
  - (A3) $\implies$ $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ **unbiased**
  - (A4) $\implies$ $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ ~~**efficient**~~ efficient among *linear* unbiased estimators

    Gauss-Markov Theorem: in the semi-parametric[4] linear regression model, we cannot show that $\hat{\beta}_{\text{OLS}}$ is efficient, but we can show that it is the most efficient among *linear*[5] unbiased estimators. It is the **Best Linear Unbiased Estimator (BLUE)**.

    $\mathbb{V}\left[\hat{\beta}|X\right] = \mathbb{E}\left[\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)\left(\hat{\beta} - \mathbb{E}\left[\hat{\beta}\right]\right)' \mid X\right] = ... = \sigma^2(X'X)^{-1}$
  - (A5) $\implies$ $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ **efficient**

    In the *parametric* linear *normal* regression model ($e_i \sim \mathcal{N}(0, \sigma^2)$), $\hat{\beta}_{\text{OLS}}$ is equal to $\hat{\beta}_{\text{MLE}}$. Therefore it is efficient, it is the **Best Unbiased Estimator (BUE)**.
  - (A4) $\implies$ $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{OLS}}$ **unbiased**[6] $\mathbb{E}\left[s^2|X\right] = \frac{1}{n-k}\mathbb{E}[r'r|X] = ... = \frac{1}{n-k}\sigma^2(n-k) = \sigma^2$
- Asymptotics

  $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ is **asymptotically unbiased** as is unbiased

  asymptotically normally distributed by a CLT, $\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, M_{\text{XX}}^{-1} M_{\text{X}\Sigma\text{X}} M_{\text{XX}}^{-1}\right)$

  asymptotically efficient as $\sigma^2(X'X)^{-1}$ is the smallest possible asymptotic variance

  consistent as 1. is asymptotically unbiased, and 2. $\mathbb{V}\left[\hat{\beta}_{\text{OLS}}|X\right] = ... \xrightarrow[n\to\infty]{p} 0$

  $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{OLS}}$ is **asymptotically unbiased** as is unbiased

  asymptotically efficient as $\sqrt{n}(\hat{\sigma}^2_{\text{OLS}} - \sigma^2) \xrightarrow{d} \mathcal{N}\left(0, 2\sigma^4\right)$

  consistent as 1. is asymptotically unbiased, and 2. $\mathbb{V}\left[\hat{\sigma}^2_{\text{OLS}}|X\right] = \frac{2\sigma^4}{n-k} \xrightarrow[n\to\infty]{p} 0$

[4] The distribution of e is not fully characterized.

[5] Here, linearity does not refer to the linearity of the model w.r.t the parameters, but to the linearity of $\hat{\beta}$ w.r.t. y, such that y enters the equation linearly: $\beta_j = \lambda_1 y_1 + ... + \lambda_n y_n$. Indeed, $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$ is linear in y.

[6] The residuals have $n-k$ degrees of freedom (k parameters $\hat{\beta}$ are estimated; the model has an intercept and $k-1$ regressors). We must hence divide by $n-k$ in order to bias-adjust any statistic that uses the residuals as proxy for the true errors.

3 Departures from the usual assumptions – and how to deal with them

3.1 OLS

3.1.1 Non-spherical errors ⟶ Sandwich estimators

Assuming (A1)–(A3), by applying the CLT, we obtain the limit distribution of the rescaled $\hat{\beta}_{\text{OLS}}$. Simply for convenience, we drop the notation $|X$, however all features of the distribution of $\hat{\beta}_{\text{OLS}}$ presented here are conditional on X:

$$\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta_0\big) = \big(\tfrac{1}{n}X'X\big)^{-1}\tfrac{1}{\sqrt{n}}X'e = \Big(\underbrace{\tfrac{1}{n}\sum_i \mathrm{x}_i\mathrm{x}_i'}_{\xrightarrow[n\to\infty]{p}\, M_{\text{XX}}}\Big)^{-1} \underbrace{\tfrac{1}{\sqrt{n}}\sum_i \mathrm{x}_i e_i}_{\xrightarrow[n\to\infty]{d}\, \mathcal{N}(0, M_{\text{X}\Sigma\text{X}})} \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, M_{\text{XX}}^{-1} M_{\text{X}\Sigma\text{X}} M_{\text{XX}}^{-1'}\Big)$$

where $M_{\text{XX}} \equiv \operatorname{plim}\big(\frac{1}{n}X'X \mid X\big) =^{7} \lim\big(\mathbb{E}\big[\frac{1}{n}X'X|X\big]\big) = \lim\big(\frac{1}{n}X'X\big)$ is finite and $\neq 0$

$M_{\text{X}\Sigma\text{X}} \equiv \operatorname{plim}\big(\frac{1}{n}X'\mathrm{e}\mathrm{e}'X \mid X\big) = \lim\big(\mathbb{E}\big[\frac{1}{n}X'\mathrm{e}\mathrm{e}'X|X\big]\big) = \lim\big(\frac{1}{n}X'\ \mathbb{E}\big[\mathrm{e}\mathrm{e}'|X\big]X\big) \equiv \lim\big(\frac{1}{n}X'\Sigma X\big)$

Σ is the variance-covariance matrix of the error term: $\mathbb{E}[\mathrm{e}\mathrm{e}'|X]$

We talk of the limit distribution of $\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta_0\big)$, instead of $\hat{\beta}_{\text{OLS}}$, because $\hat{\beta}_{\text{OLS}}$ has a degenerate distribution with all mass at β. However, it would be more convenient to think of the distribution of $\hat{\beta}_{\text{OLS}}$ rather than $\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta_0\big)$. We do this by introducing the artifice of "asymptotic distribution". We consider n large but not infinite, s.t. the asymptotics have kicked in, then we can drop the limits in the expressions (lim is dropped, plim becomes $\mathbb{E}$). We obtain $\hat{\beta}_{\text{OLS}}$'s asymptotic distribution:

$$\hat{\beta}_{\text{OLS}} \overset{\text{a}}{\sim} \mathcal{N}\Big(\beta_0, \tfrac{1}{n}\big(\tfrac{1}{n}X'X\big)^{-1}\big(\tfrac{1}{n}X'\Sigma X\big)\big(\tfrac{1}{n}X'X\big)^{-1'}\Big) = \mathcal{N}\Big(\beta_0, \underbrace{\big(X'X\big)^{-1}\ X'\Sigma X\ \big(X'X\big)^{-1'}}_{\overset{\text{a}}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]}\Big)$$

We need a **consistent** estimate of the asymptotic variance-covariance matrix $\overset{\text{a}}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]$ in order to do (sampling-based) statistical inference[8]. One approach is to use **sandwich estimators**[9]. The only unknown is Σ. We decompose the variance into its 3 $k \times k$ components: *bread, meat, bread*, and select a **consistent** estimator of the *meat* component that best represents our assumed error structure, to finally compute:

$$\widehat{\overset{\text{a}}{\mathbb{V}}}\big[\hat{\beta}_{\text{OLS}}\big] \equiv \underbrace{\big(X'X\big)^{-1}}_{bread}\ \underbrace{X'\widehat{\Sigma}X}_{meat}\ \underbrace{\big(X'X\big)^{-1'}}_{bread}$$

Error structure

[7] For a sample average $\bar{Z}_n$: by an LLN, plim $\bar{Z}_n = \lim \mathbb{E}[\bar{Z}_n]$.

[8] The standard errors used in the t-test for $\hat{\beta}_{\text{OLS}}$ are none other than an estimate for $\sqrt{\overset{\text{a}}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]}$.

[9] All extremum estimators can actually be shown to be consistent and asymptotic normal, with an asymptotic variance matrix in the sandwich form: $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, A(\theta)^{-1}B(\theta)A(\theta)^{-1}\big)$. The sandwich algorithm presented here for OLS can be extended to all extremum estimators, e.g., MLE and GMM. Which is not to say one should. Freedman (2006) points out that while White's sandwich estimator often gives good results in OLS, the equivalent correction in MLE does not necessarily make sense: *"If the model is nearly correct, so are the usual standard errors, and robustification is unlikely to help much. On the other hand, if the model is seriously in error, the sandwich may help on the variance side, but the parameters being estimated by the MLE are likely to be meaningless."* (If the specification–and hence the likelihood function–is incorrect, the parameter will be biased; why care about the variance of an estimator for the wrong parameter.)

Spherical (A4)

$\Sigma = \sigma^2 \mathrm{I}$, therefore $\overset{\text{a}}{\mathbb{V}} = \left(X'X\right)^{-1} X'\Sigma X \left(X'X\right)^{-1'} = \sigma^2 \left(X'X\right)^{-1} (X'X) \left(X'X\right)^{-1'} = \sigma^2 \left(X'X\right)^{-1}$

We can consistently estimate the population variance σ^2 by the unbiased sample variance $s^2 = \frac{\sum_i r_i^2}{n-k}$, and hence $\overset{\text{a}}{\mathbb{V}}$ by $\widehat{\overset{\text{a}}{\mathbb{V}}}_{\text{S}} \equiv s^2 \left(X'X\right)^{-1}$.

This expression is actually the Cramer-Rao lower bound, therefore $\hat{\beta}_{\text{OLS}}$ is **BLUE.**

Not spherical

- **Heteroskedastic**

  White (1980) proposes a non-parametric estimator which produces heteroskedasticity-consistent (HC) or "robust" standard errors. They will be larger than those assuming homoskedasticity (which are downward-biased), as they account for the extra variation. HC SEs seem to have become best practice with large samples, as one can rarely assume homoskedastic errors[10].

  Use $\widehat{\Sigma}_{\text{H}} = \frac{1}{n-k} diag[r_i^2]$, i.e., $X'\widehat{\Sigma}_{\text{H}} X = \frac{1}{n-k} \sum_i r_i^2 \mathrm{x}_i \mathrm{x}_i'$.

  The resulting $\widehat{\overset{\text{a}}{\mathbb{V}}}_{\text{H}}$ is consistent for $\overset{\text{a}}{\mathbb{V}}$, even though r_i^2 is not consistent for σ_i^2.

- **Autocorrelated**

  If errors are autocorrelated in any way (in time, space, both, by groups...), it means that our model is not capturing some feature of the DGP. There are essentially two ways to deal with this structure, in order to conduct proper inference:

  1. Treat it as *substance*: incorporate the structure in the model. This also enables us to *study* it. *Ex: if errors are autocorrelated by group, model a multilevel data structure.*
  2. Treat it as *nuisance*: not incorporate the structure, but adjust for it after fitting the model[11]. *Ex: if errors are autocorrelated by group, cluster the standard errors.*

  The sandwich estimators below correspond to the second approach. These are also built with the same strategy:

  - *There is autocorrelation in e_i, in some dimension(s) (time, space, group...) in the population.*
  - *The true covariance matrix of the errors Σ, and therefore the true asymptotic covariance matrix of $\hat{\beta}_{\text{OLS}}$, will contain these non-zero non-diagonal terms. We want our estimate of $\overset{\text{a}}{\mathbb{V}}$ to be consistent, therefore we must produce a $\widehat{\Sigma}$ that estimates these terms consistently.*
  - *Two assumptions are made: 1. the process is **2nd order stationary**[12], 2. which covariance terms are potentially non-zero and the weights we give them (e.g., by using a kernel function).*
  - *We estimate each of these terms by their sample equivalent. I.e., $\widehat{\Sigma}$ is made of diagonal terms that are the same as White's (these estimators are hence also heteroskedasticity-consistent), and some non-zero non-diagonal terms that are the **sample autocovariances**.*

[10] Ideally, we would calculate an efficient estimator directly, instead of accepting an inefficient OLS and adjusting the SEs. The appropriate estimator is weighted least squares (WLS). However, its asymptotic efficiency rests on the correct specification of the pattern of heteroskedasticity. I.e., WLS is the better solution if we know the pattern, but we usually don't.

[11] If we make no adjustments for this structure, default standard errors will greatly overstate estimator precision. Note that similarly as the note above, the first-best strategy would be to use generalized least squares (GLS), which produce an efficient estimator if we know the correct specification of the pattern of autocorrelation; but we usually don't.

[12] A stationary process is "a stochastic process whose unconditional joint probability distribution does not change over the dimension of the process". I.e., here:

· for a time series: the autocorrelation between 2 obs. that are m periods apart, is the same across the period;
· for a spatial process: the autocorrelation between 2 obs. that are apart by a distance d, is the same across the spatial field;
· for a process across groups: the autocorrelation between 2 obs. is fully determined by their group appartenance.

∗ **Serial (temporal) correlation**

The dimension along which autocorrelation occurs is time.

Newey and West (1987) propose an estimator that accounts for serial correlation of unknown form in the errors of a single time series. It can be expanded to panel datasets, by estimating only correlations between lagged errors in the same cluster.

Consider a single time series $\{e_t\}$, and:
- Its autocovariance of lag l: $\gamma_e[t, t-l] \equiv cov[e_t, e_{t-l}] = \mathbb{E}[(X_t - \mu_t)(X_{t-l} - \mu_t)]$.
  If the process is covariance-stationary, it is a function of the relative lag only: $\gamma_e[l]$
- Its bias-adjusted sample equivalent, for a sample $i = 1, ..., T$: $g(l) = \frac{1}{T-k} \sum_{t=l+1}^{T} r_t r_{t-l}$.

The Newey-West estimator weights these covariance estimates using a triangular kernel function[13], s.t. the weight decreases linearly with the lag up to a chosen maximum lag L, and adds White's variance estimates:

$$\widehat{\Sigma}_{\text{NW}} = G(0) + \sum_{l=1}^{L} \left(1 - \tfrac{l}{L+1}\right) [G(l) + G(l)']$$

$$\text{Use } X'\widehat{\Sigma}_{\text{NW}}X = \tfrac{1}{T-k} \sum_{t=1}^{T} r_t^2 \mathrm{x}_t \mathrm{x}_t' + \sum_{l=1}^{L} \left(1 - \tfrac{l}{L+1}\right) \left[\tfrac{1}{T-k} \sum_{t=l+1}^{n} r_t r_{t-l} \mathrm{x}_t \mathrm{x}_{t-l}' + \tfrac{1}{T-k} \sum_{t=l+1}^{n} r_t r_{t-l} \mathrm{x}_{t-l} \mathrm{x}_t' \right]$$

$$= \tfrac{1}{T-k} \sum_{t=1}^{T} r_t^2 \mathrm{x}_t \mathrm{x}_t' + \tfrac{1}{T-k} \sum_{l=1}^{L} \left(1 - \tfrac{l}{L+1}\right) \sum_{t=l+1}^{n} r_t r_{t-l} (\mathrm{x}_t \mathrm{x}_{t-l}' + \mathrm{x}_{t-l} \mathrm{x}_t')$$

Notes:
- $\widehat{\Sigma}$ *is consistent iff* $L \to \infty$ *and* $\frac{L}{T^{1/4}} \to 0$ *as* $T \to \infty$*, i.e., iff* L *grows slower than* $T^{1/4}$*. A common practice is hence to set* L *to the integer part of* $T^{1/4}$*.*

∗ **Spatial correlation**

The dimension along which autocorrelation occurs is space.

This dimension is actually a dual dimension: while time or group appartenance are 1D, space is at least 2D. Conley (1999), under the supplementary assumption that the process is isotropic, proposes a consistent estimator for $\widehat{\Sigma}$ that accounts for spatial correlation of unknown form in the errors, and for heteroskedasticity. It weights the sample covariances using a kernel function $k(s_i, s_j)$, where s_i is the location of observation i.

$$X'\widehat{\Sigma}X = \tfrac{1}{n-k} \sum_{i=1}^{n} r_i^2 \mathrm{x}_i \mathrm{x}_i' + \tfrac{1}{n-k} \sum_{i=1}^{n} \sum_{j=1}^{n} k(s_i, s_j)\ e_i e_j \mathrm{x}_i \mathrm{x}_j'$$

Notes:
- *Multiple choices of kernel are possible. Conley (2008) presents the uniform kernel but does not recommend it over another.* $\widehat{\Sigma}$ *will be consistent if* $\forall h, k(s, s+h) \to 1$ *as* $n \to \infty$*, but slowly enough for the variance of* $\widehat{\Sigma}$ *to collapse to zero. Assuming a stationary and isotropic process,*

[13] The modified Bartlett weights also ensure that $\widehat{\Sigma}$ is positive semi-definite, which is required for the formation of asymptotic confidence interval and hypothesis testing.

$k(i,j)$ simplifies to a function of distance: $k(d_{ij})$. Choose whichever distance metric fits the context, e.g., a metric of economic distance.
- *Conley (1999) shows that spatial dependence does not imply that SEs will necessarily rise. In his empirical example, 6 of 9 spatial SE estimates are smaller than their iid counterparts.*
- *This estimator is very similar to the method of Kriging in geostatistics.*

∗ **Clustering**

The dimension along which autocorrelation occurs is group appartenance.

Errors are correlated within (but not between) groups or "clusters". I.e., we think the DGP is: $y_{i,g[i]} = \mathrm{x}'_{i,g[i]}\beta + e_{i,g[i]}$, where $\mathbb{E}[e_i|\mathrm{x}_i] = 0$, $\mathbb{E}[e_i e_j|\mathrm{x}_i, \mathrm{x}_j] = 0$ only if $i, j \notin$ same g. The covariance matrix of the error term Σ has a block-diagonal structure.

Use $X'\widehat{\Sigma}X = \frac{1}{n-k}\sum_{g=1}^{G} X'_g \mathrm{r}_g \mathrm{r}'_g X_g$. The resulting *cluster-robust* $\overset{\widehat{\mathrm{a}}}{\mathbb{V}}_{\mathrm{C}}$ is consistent for $\overset{\mathrm{a}}{\mathbb{V}}$.

Notes:
- *This method is fully non-parametric, it allows for arbitrary dependence within a cluster.*
- $\overset{\widehat{a}}{\mathbb{V}}_{\mathrm{C}}$ *is also heteroskedasticity-consistent. It is typically* $> \overset{\widehat{a}}{\mathbb{V}}_{\mathrm{H}}$ *due to the addition of all non-diagonal terms within clusters.*
- ⚠ *Do not use this estimator with too few clusters (rule of thumb: have* > 40*), as, like all sandwich estimators, it relies on asymptotics*[14]*. Cameron and Miller (2015) recommend at least using critical values from the* t_{G-1} *distribution instead of the normal* $\mathcal{N}(0,1)$*. Note also that if clusters are unbalanced, the effective number of clusters is actually even lower.*

[14] The t-statistic $t_{\hat{\beta}} = \dfrac{\hat{\beta} - \beta_0}{\sqrt{\widehat{\mathbb{V}}_{\mathrm{C}}[\hat{\beta}]}} \underset{h_0}{\overset{a}{\sim}} \mathcal{N}(0,1)$, however for finite G (and therefore, especially for small G), the statistic's distribution is unknown – even with normal errors. Using critical values from the standard normal distribution will downward-bias the variance estimate – leading to too narrow confidence intervals and over-rejection of the null.

A Misc.

A.1 Linear algebra — Positive-definite matrices

An $k \times k$ matrix A is **invertible** if there exists an $k \times k$ matrix B such that $AB = BA = I_k$.
A square matrix that is not invertible is called **singular or degenerate**.

The quasi-totality of square matrices are invertible.

Let M be an $k \times k$ symmetric real matrix, $\{\lambda_k\}$ its eigenvalues.

- M is **positive-definite** $\iff z'Mz > 0$ for every vector $z \in \mathbb{R}^k \iff$ all $\{\lambda_k\}$ are > 0.
- M is **positive semi-definite** $\iff z'Mz \geq 0$ for every vector $z \in \mathbb{R}^k \iff$ all $\{\lambda_k\}$ are ≥ 0.

Ex: The identity matrix I_k is positive-definite.

– Every positive definite matrix is invertible and its inverse is also positive definite.
– In statistics, the covariance matrix of a multivariate probability distribution is always symmetric and positive semi-definite; and it is positive definite unless one variable is an exact linear function of the others. Conversely, every positive semi-definite matrix is the covariance matrix of some multivariate distribution. Here we are talking about *population* covariance matrices. It is possible that the *sample* covariance matrix is singular, e.g., if there is exact collinearity, or when the number of observations is less than the number of variables.

A.2 Kernel functions for non-parametric statistics

A **kernel** is a non-negative real-valued integrable function $k()$, used as a **weighting function** in non-parametric estimation techniques. They are also called "window functions" (notably in time-series).

Some applications require the function to satisfy additional conditions, for instance:

– normalization: $\int_{-\infty}^{+\infty} k(u)\,du = 1$
In kernel density estimation, this ensures that the estimation produces a probability density function.

– symmetry: $\forall u,\ k(-u) = k(u)$
This ensures that the average of the corresponding distribution is equal to that of the sample used.

Examples of commonly used symmetric kernels, with the arbitrary bounded support $[-1, 1]$[15], i.e., $|u| \leqslant 1$:

- uniform: $k(u) = 1$ for $|u| \leqslant 1$
- triangular (Bartlett): $k(u) = 1 - |u|$ over the support $|u| \leqslant 1$
- parabolic (Epanechnikov): $k(u) = \frac{3}{4}(1 - u^2)$ over the support $|u| \leqslant 1$

[15] S.t. $k(u) = 0$ for values of u outside the support.

References

Cameron, A. C. and Miller, D. L. A practitioner's guide to cluster-robust inference. The Journal of Human Resources, 50(2):317–372, 2015. ISSN 0022-166X.

Conley, T. GMM estimation with cross sectional dependence. Journal of Econometrics, 92(1):1–45, 1999. ISSN 0304-4076. doi: 10.1016/S0304-4076(98)00084-0.

Conley, T. G. Spatial Econometrics. In Durlauf, S. and Blume, L., editors, The New Palgrave Dictionary of Economics, volume 7, pages 741–747. 2nd edition, 2008. doi: 10.1057/978-1-349-95121-5_2023-1.

Freedman, D. A. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". The American Statistician, 60(4):299–302, 2006. doi: 10.1198/000313006X152207.

Gelman, A., Hill, J., and Vehtari, A. Regression and Other Stories. Cambridge University Press, July 2020. ISBN 978-1-107-02398-7. doi: 10.1017/9781139161879.

Newey, W. K. and West, K. D. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica, 55(3):703–708, 1987. doi: 10.2307/1913610.

White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica, 48(4):817–838, 1980. doi: 10.2307/1912934.