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).

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	$\mathbb{E}[e_i \mid \mathbf{X}] = 0$	$\mathbb{E}[\mathrm{e} \mid \mathbf{X}] = 0_{N \times 1}$
A4. spherical errors	$e_i \mid \mathbf{X} \overset{\text{iid}}{\sim} (0, \sigma^2)$	$\mathbb{V}[\mathrm{e} \mid \mathbf{X}] = \sigma^2 \mathbf{I}_N$
– **independent errors**	$\mathrm{cov}[e_i, e_j \mid \mathbf{X}] = 0$	
– **homoskedastic errors**	$\mathbb{V}[e_i \mid \mathbf{X}] = \sigma^2 \cancel{\sigma_i^2}$	
A5. normal errors	$e_i \mid \mathbf{X} \sim \mathcal{N}(0, \sigma^2)$	$\mathrm{e} \mid \mathbf{X} \sim \mathcal{N}(0_{N \times 1}, \sigma^2 \mathbf{I}_N)$

(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.
$\mathbb{E}[\mathrm{e} \mid \mathbf{X}] = 0$ 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: $\mathrm{cov}(e_i, e_j \mid X) = \mathbb{E}[e_i e_j \mid X] = 0$
  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 $\mathbb{V}[e_i \mid X] = \sigma^2$
  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 $\beta = 0$, assumes that the sampling distribution of $\hat{\beta}$ is normal. If errors are not normal, then $\hat{\beta}$ isn't normal either.
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. t and F tests are hence robust to departures from normality, when n is large enough.
Yet, 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).

2 OLS and ML estimators of $\boldsymbol{\theta} = \left\{\boldsymbol{\beta}, \boldsymbol{\sigma^2}\right\}$ 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, since we always deal with finite samples, the behaviour of estimators in finite samples may seem more important. In effect, bias and efficiency are the two most commonly used 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 start from different motivations and are rooted in very different mathematical disciplines: calculus for OLS, and probabilities for ML.
- OLS is often used, because $\hat{\beta}_{\text{OLS}}$ is unbiased even in the case of 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 $\hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{MLE}}$.

2.2 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:

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

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[1]:

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

Properties [assuming (A1)–(A5)]

- Finite samples

$\hat{\boldsymbol{\beta}}_{\text{MLE}}$ 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 $\mathbb{V}\big[\hat{\beta}_{\text{MLE}}|X\big] = \mathbb{E}\Big[\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)' \mid X\Big] = \mathbb{E}\Big[\big(\hat{\beta} - \beta_0\big)\big(\hat{\beta} - \beta_0\big)' \mid X\Big]$

$$= \mathbb{E}\big[(X'X)^{-1}X'e\big((X'X)^{-1}X'e\big)'|X\big]$$

$$= (X'X)^{-1}X'\ \mathbb{E}[ee'|X]\ X(X'X)^{-1} = \sigma^2(X'X)^{-1} \leqslant \mathbb{V}\big[\hat{\beta}_{...}|X\big]$$

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

$\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\text{MLE}}$ 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

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

asymptotically efficient as is efficient

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

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

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

[1] 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 $\hat{\boldsymbol{\theta}}_{\text{OLS}} = \left\{\hat{\boldsymbol{\beta}}_{\text{OLS}}, \hat{\boldsymbol{\sigma}}^2_{\text{OLS}}\right\}$

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.

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

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) $\Longrightarrow$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ **unbiased**

(A4) $\Longrightarrow$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ ~~**efficient**~~ efficient among *linear* unbiased estimators

Gauss-Markov Theorem: in the semi-parametric[2] 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*[3] unbiased estimators. It is the **Best Linear Unbiased Estimator (BLUE)**.

$$\mathbb{V}\big[\hat{\beta}|X\big] = \mathbb{E}\left[\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)' \mid X\right] = ... = \sigma^2(X'X)^{-1}$$

(A5) $\Longrightarrow$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ **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) $\Longrightarrow$ $\hat{\boldsymbol{\sigma}}^2_{\text{OLS}}$ **unbiased**[4] $\mathbb{E}\big[s^2|X\big] = \frac{1}{n-k}\mathbb{E}[r'r|X] = ... = \frac{1}{n-k}\sigma^2(n-k) = \sigma^2$

- Asymptotics

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

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

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

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

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

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

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

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

[3] 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.

[4] 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 $\longrightarrow$ 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:[5]

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

which leads to $\hat{\beta}_{\text{OLS}}$'s "asymptotic distribution": $\hat{\beta}_{\text{OLS}} \overset{a}{\sim} \mathcal{N}\Big(\beta_0, \underbrace{\tfrac{1}{n}M_{\text{XX}}^{-1}M_{\text{X}\Sigma\text{X}}M_{\text{XX}}^{-1}}_{\overset{a}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]}\Big)$

where $M_{\text{XX}} \equiv \text{plim}\big(\tfrac{1}{n}X'X\big) = \text{plim}\Big(\tfrac{1}{n}\sum_i \mathrm{x}_i\mathrm{x}_i'\Big) =^{6} \lim\Big(\tfrac{1}{n}\sum_i \mathbb{E}[\mathrm{x}_i\mathrm{x}_i']\Big)$ is finite and $\neq 0$;

$M_{\text{X}\Sigma\text{X}} \equiv \text{plim}\big(\tfrac{1}{n}X'\mathrm{ee}'X\big) = \text{plim}\Big(\tfrac{1}{n}\sum_i e_i^2\mathrm{x}_i\mathrm{x}_i'\Big) = \lim\Big(\tfrac{1}{n}\sum_i \mathbb{E}[e_i^2\mathrm{x}_i\mathrm{x}_i']\Big)$;

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

We need a consistent estimate of the asymptotic variance-covariance matrix $\overset{a}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]$ in order to do (sampling-based) statistical inference[7]. One approach is to use **sandwich estimators**. We decompose the variance into its 3 $k \times k$ components: *bread, meat, bread*, and compute a **consistent** estimator of each:

$$\overset{a}{\mathbb{V}}\big[\hat{\beta}_{\text{OLS}}\big] = \tfrac{1}{n}\ \underbrace{M_{\text{XX}}^{-1}}_{bread}\ \underbrace{M_{\text{X}\Sigma\text{X}}}_{meat}\ \underbrace{M_{\text{XX}}^{-1}}_{bread}$$

The bread M_{XX} can be consistently estimated by $\frac{1}{n}X'X$, by definition. The focus is on the meat: we need to select a consistent $\widehat{\Sigma}$ that best represents our assumed[8] error structure, to finally compute:

$$\widehat{\overset{a}{\mathbb{V}}}\big[\hat{\beta}_{\text{OLS}}\big] = \tfrac{1}{n}\ \big(\tfrac{1}{n}X'X\big)^{\text{-}1}\tfrac{1}{n}X'\widehat{\Sigma}X\big(\tfrac{1}{n}X'X\big)^{\text{-}1} = (X'X)^{\text{-}1}\ X'\widehat{\Sigma}X\ (X'X)^{\text{-}1}$$

Error structure

[5] Note that the formulas in this section combine matrices, vectors and scalars. Notably: $X'X$ is a $k \times k$ matrix, which we also write in the form of $\sum_i \mathrm{x}_i\mathrm{x}_i'$ as x_i is a $k \times 1$ vector. $X'e$ is itself a $k \times 1$ vector, which we also write in the form of $\sum_i \mathrm{x}_i e_i$, where e is the $n \times 1$ vector of error terms, and e_i is a scalar. The final asymptotic variance matrix is $n \times n$.

[6] By an LLN, plim $\bar{Z}_n = \lim \mathbb{E}[\bar{Z}_n]$. Here we have the sample average $\bar{Z}_n \equiv \frac{1}{n}\sum_i x_i x_i' \xrightarrow[n\to\infty]{p} \mathbb{E}[\frac{1}{n}\sum_i x_i x_i'] = \frac{1}{n}\sum_i \mathbb{E}[x_i x_i']$.

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

[8] This is not to say that one should make assumptions about e ex-ante, and not check them after. Plotting the residuals of a first model that is agnostic w.r.t. the error term (as one should do after fitting *any* model, if the estimation method makes assumptions about the error term), can reveal a pattern one might have not foreseen. We can then use that pattern to formulate an appropriate assumption about the structure of e, and compute a consistent $\widehat{\Sigma}$.

Spherical: (A4) $\boldsymbol{\Sigma = \sigma^2 I}$

$$M_{\text{X}\Sigma\text{X}} \equiv \text{plim}\big(\tfrac{1}{n}X'\text{ee}'X\big) = \lim \mathbb{E}[\tfrac{1}{n}X'\text{ee}'X|X] = \text{plim}\big(\tfrac{1}{n}X'\mathbb{E}[\text{ee}'|X]X\big) = \sigma^2 \;\text{plim}\big(\tfrac{1}{n}X'X\big)$$

$$\implies \overset{\text{a}}{\mathbb{V}} = \tfrac{1}{n}M_{\text{XX}}^{-1}M_{\text{X}\Sigma\text{X}}M_{\text{XX}}^{-1} = \tfrac{1}{n}\;\text{plim}\big(\tfrac{1}{n}X'X\big)^{\text{-1}}\;\sigma^2\text{plim}\big(\tfrac{1}{n}X'X\big)\;\text{plim}\big(\tfrac{1}{n}X'X\big)^{\text{-1}} = \tfrac{1}{n}\sigma^2\;\text{plim}\big(\tfrac{1}{n}X'X\big)^{\text{-1}}$$

We can consistently estimate:

$$\left.\begin{array}{l} \text{– the pop. variance } \sigma^2 \text{ by the bias-adjusted sample variance } s^2{=}\frac{\sum_i r_i^2}{n-k} \\ \text{– plim}\big(\tfrac{1}{n}X'X\big) \text{ by } \tfrac{1}{n}X'X \end{array}\right\} \implies \overset{\text{a}}{\mathbb{V}} \text{ by } \overset{\widehat{\text{a}}}{\mathbb{V}} \equiv s^2\big(X'X\big)^{\text{-1}}$$

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

Not spherical

- **Heteroskedastic**

  We can compute heteroskedasticity-consistent (HC) or "robust" standard errors, following White (1980). 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[9].

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

  The resulting $\overset{\widehat{\text{a}}}{\mathbb{V}}_{\text{H}}$ 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 time, in space, 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 feature of the data in our model – enabling us to also study it (ex: if errors are autocorrelated by group, by modeling a multilevel data structure);
  2. treat it as *nuisance*: not incorporate but correct for this structure after fitting the model[10]. Otherwise, default standard errors would greatly overstate estimator precision.

  The sandwich estimators below correspond to this second approach.

  - **Serial (temporal) correlation**

    Newey and West (1987) propose a procedure to account for serial correlation of unknown form in the residuals of a single time series. It can be modified for use in a panel data set, by estimating only correlations between lagged residuals in the same cluster.

    In the context of a single time series: we want to estimate $\frac{1}{T}X'\Sigma X = \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\rho_{|t-s|}X_tX_s'$, where $\rho_{|t-s|}$ is the serial correlation between 2 observations that are $t-s$ periods apart.

    The Newey-West estimator uses a weighting function: the covariance term of lag l (e.g., e_te_{t-l}) is multiplied by the weight $1-\frac{l}{L+1}$ which decreases with the time passed between the two disturbances, where L is the chosen maximum lag.

[9] Ideally, we would calculate an efficient estimator directly, instead of accepting an inefficient OLS and adjusting the standard errors. The appropriate estimator is weighted least squares (WLS). However, its asymptotic efficiency depends 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.

[10] Similarly, generalized least squares (GLS) would produce an efficient estimator, provided that we know the correct specification of the pattern of autocorrelation; and we usually don't.

We compute the L-lag consistent estimator:

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

- **Spatial correlation**

- **Clustering**

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

Notes:

· *This method is non-parametric, it allows for arbitrary dependence within a cluster.*

· $\overset{\widehat{a}}{\mathbb{V}}_{\mathrm{C}}$ *is actually both cluster- and 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*[11]*. Cameron and Miller (2015) recommend at least using critical values from the* $\mathcal{T}_{G-1}$ *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.*

[11] The t-statistic $t_{\hat{\beta}} = \dfrac{\hat{\beta}-\beta_0}{\sqrt{\hat{\mathbb{V}}_{\mathrm{C}}[\hat{\beta}]}} \overset{a}{\underset{h_0}{\sim}} \mathcal{N}(0,1)$, 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.

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. URL http://www.jstor.org/stable/24735989.

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.