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

Contents

Disclaimer: Sections and lines in brown are 'under construction'.

1 Assumptions of the CLRM for inference

The classical linear regression model (CLRM) consists of a set of *population* assumptions that describe the data generating process (DGP). By decreasing order of importance (Gelman et al., 2020, Ch. 11):

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

(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 collinearity).
If this is violated, drop one regressor, or transform collinear regressors into a single x.

(A3) Strict[1] **exogeneity of regressors**: all other factors that affect y_i are unrelated to x_i.
$\mathbb{E}[e_i|\mathrm{X}] = 0$ *also implies* $\mathbb{E}[e_i]$=0 *and* $\mathbb{E}[\mathrm{X}'e_i]$=0*, leading to cov*$[e_i, \mathrm{x}] = 0$*:* X *and* e_i *are uncorrelated.*

(A4) Spherical errors[2]

- **Independent errors**: errors are randomly spread around the regression line.
  $\implies$ no autocorrelation: $\mathrm{cov}[e_i, e_j|\mathrm{X}] = \mathbb{E}[e_i e_j|\mathrm{X}] = 0$
  This will be violated when there is structure in the data that is left out of the model (e.g., with time series data, there may be serial correlation in the error term if the model doesn't include lags of regressors, nor is an autoregressive or a moving average model...).
- **Homoskedastic errors**: equal *conditional* variance $\mathbb{V}[e_i|\mathrm{X}] = \sigma^2$
  I.e., the spread of errors, or model uncertainty, is identical across the support of y_i.
  If this is violated, $\hat{\beta}_{\text{OLS}}$ remains valid but is inefficient: $\hat{\beta}_{\text{WLS}}$ has lower variance.

(A5) Normal errors
This assumption is not required for estimating the regression but for making **inferences**, e.g., computing confidence intervals or p-values. Without (A5), t and F tests are invalid.
To perform statistical inference, we need to know the full sampling distribution of $\hat{\beta}$ (ex: the one-sample t-test of $H_0 : \beta$=0 assumes that the sampling distribution of $\hat{\beta}$ is normal). This sampling distribution depends on the distribution of e_i; if errors aren't normal, then $\hat{\beta}_{\text{OLS}}$ 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 $\hat{\beta}_{\text{OLS}}$ is normal, such that t and F tests are robust to departures

[1] "Strict" exogeneity refers to the fact that the expected value of e_i is not related to $\mathrm{x}_j, \forall j$ (not solely x_i). Ex: with time series data, this means not only that e_t and x_t are uncorrelated, but that e_t is uncorrelated with past and future values of X.

[2] Note that most statements are actually conditional statements. E.g., (A4) assumes *conditionally* homoskedastic errors.

from normality if n is large. I.e., with large sample sizes, non-normality of e_i isn't a big problem as the normality of $\widehat{\beta}_{\text{OLS}}$ is still approximately true. But with small n and highly non-normal errors, appealing to an asymptotically normal approximation may be unreasonable, and one may want to consider an alternative (e.g., bootstrap).

2 Estimators of $\{\beta, \sigma^2\}$ and their statistical properties

Multiple estimators are often available to estimate some summary of the relationship between x_i and y_i. How one chooses between them (besides their ease of computation) is motivated by their **statistical properties.**

2.1 Estimator properties

Let $\hat{\theta}$ be an estimator for the population parameter θ, for a sample of size n. $\hat{\theta}$, as a function of the random sample, is a random variable. The various possible samples of size n would each lead to a different realization $\hat{\theta}_s$, which together form the estimator's density or probability distribution function (pdf) $f_{\hat{\theta}}$. $\hat{\theta}$ has:

— finite sample properties: characteristics of $f_{\hat{\theta}}$ for a finite n. *Ex: bias, efficiency (precision);*
— asymptotic properties: characteristics of $f_{\hat{\theta}}$ as $n \to \infty$. *Ex: consistency, asymptotic distribution.*

As we always deal with finite samples, finite sample properties may seem the most important. In effect, bias (concerned with the center of the pdf) and efficiency (its spread) are the most common selection criteria. We note however that these tell us nothing about the properties of the estimator *for our own sample,* they tell us only about the distribution of values from hypothetical samples.

Finite sample properties

- $\hat{\theta}$ is **unbiased** iff $\mathbb{E}\big[\hat{\boldsymbol{\theta}}\big] = \boldsymbol{\theta}$
  The estimator is correct *in expectation* over all possible samples. I.e., its distribution is centered around the estimand. ⚠ *But our estimate from our own sample could be anywhere within that distribution, e.g., far from its center.* The bias of $\hat{\theta}$ is $\mathbb{E}[\hat{\theta}] - \theta$.
- $\hat{\theta}$ is **efficient** or "best" iff it has the lowest possible variance of all estimators: $\mathbb{V}\big[\hat{\boldsymbol{\theta}}\big] \leqslant \mathbb{V}\big[\tilde{\boldsymbol{\theta}}_{...}\big]$
  Its distribution is condensed, thus though the realization $\hat{\theta}_s$ from any sample could be anywhere within that distribution, it will never be too far away from the mean (which, if the estimator is unbiased, is the true θ). For unbiased estimators, that variance is the Cramér-Rao lower bound.

Asymptotic properties

- $\hat{\theta}$ is **asymptotically unbiased** iff $\mathbb{E}\big[\hat{\boldsymbol{\theta}}\big] \xrightarrow[n\to+\infty]{p} \boldsymbol{\theta}$
- $\hat{\theta}$ is **asymptotically efficient** iff $\mathbb{V}\big[\hat{\boldsymbol{\theta}}\big] \xrightarrow[n\to+\infty]{p}$ ***asymptotic Cramér-Rao lower bound***
- $\hat{\theta}$ is **consistent** iff $\hat{\boldsymbol{\theta}} \xrightarrow[n\to+\infty]{p} \boldsymbol{\theta}$
  I.e., as we get enough data, then we know the truth. Sufficient conditions are that (i) $\hat{\theta}$ be asymptotically unbiased, and (ii) its variance $\to 0$ as $n \to \infty$.

In frequentist statistics, the Maximum Likelihood (ML) and Ordinary Least Squares (OLS) estimators are widely used. The next sections describe them and their properties.[3] Preliminary remarks:

— OLS and ML are rooted in different mathematical disciplines: probabilities for ML, calculus for OLS (OLS makes no assumption on the probabilistic nature of the variables, it is deterministic).
— OLS is tailored to the linear regression model, and is often used because the estimated function $f(\mathrm{x}_i, \hat{\beta}_{\text{OLS}})$ is unbiased for the conditional expectation $\mathbb{E}[y_i|\mathrm{x}_i]$ even with non-spherical errors.
— ML includes OLS as a special case: if $e_i|\mathrm{x}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, then $y_i|\mathrm{x}_i \sim \text{MVN}(\mathrm{x}_i'\beta, \sigma^2)$ and $\hat{\beta}_{\text{OLS}} = \hat{\beta}_{\text{ML}}$.

[3] All features presented are of the estimators' *conditional* distributions, we simply drop the notation '$|\mathrm{x}_i$' for convenience.

2.2 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 $y = g(\mathrm{x}, \beta)$ to each data point is measured by its residual $r_i \coloneqq y_i - g(x_i, \beta)$. The Ordinary Least Squares (OLS) estimator computes, in the context of a model linear in the parameters $g(\mathrm{x}, \beta) = \sum_{j=1}^{p} \beta_j h_j(\mathrm{x})$, the values of the parameters that minimize the sum of the squares of the residuals:

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

Solution Let X be the matrix of transformed regressors $\{h_j(\mathrm{x})\}_{j=1}^{p}$. The FOC of the minimization problem gives an exact closed-form solution (which the SOC guarantees is a minimum iff the matrix $\mathrm{X}'\mathrm{X}$ is positive definite):

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

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

Properties [assuming (A1)-(A3)]

- Finite samples

  (A3) $\implies$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ **unbiased**

  (A4) $\implies$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ ~~**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)**.

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

  (A5) $\implies$ $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ **efficient**
  In the *parametric* linear *normal* regression model ($e_i \sim \mathcal{N}(0, \sigma^2)$), $\hat{\beta}_{\text{OLS}}$ is equal to $\hat{\beta}_{\text{ML}}$. Therefore it is efficient, it is the **Best Unbiased Estimator (BUE)**. It is also normally distributed: $\hat{\beta}_{\text{OLS}} = \beta + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{e} \sim \mathcal{N}\left(\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\right)$.

  (A4) $\implies$ $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\text{OLS}}$ **unbiased**[6] $\mathbb{E}\left[s^2|\mathrm{X}\right] = \frac{1}{n-p}\mathbb{E}[\mathrm{r}'\mathrm{r}|\mathrm{X}] = \ldots = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$

- Asymptotics

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

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

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

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

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

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

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

[4] The distribution of e_i 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. the vector y, such that y enters the equation linearly: $\beta_j = \lambda_1 y_1 + \ldots + \lambda_n y_n$. Indeed, $\hat{\beta}_{\text{OLS}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$ is linear in y.

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

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

Definition The likelihood function in a regression model is the probability density of the data given the parameters θ and predictors. The Maximum Likelihood (ML) estimator is then the value of the parameters θ s.t. under the assumed model, the observed data are most likely. Assuming iid observations, we have:

- the likelihood $\mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big) = f\big(\mathrm{x}_1,...,\mathrm{x}_n,\theta\big) \overset{iid}{=} f\big(\mathrm{x}_1,\theta\big) \times ... \times f\big(\mathrm{x}_n,\theta\big) = \prod_{i=1}^{n} f\big(\mathrm{x}_i,\theta\big)$
- the log-likelihood $\log \mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big) = \sum_{i=1}^{n} \log f\big(\mathrm{x}_i,\theta\big)$
- the ML estimator $\hat{\theta}_{\text{ML}} := \underset{\theta}{\text{argmax}}\ \mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big) = \underset{\theta}{\text{argmax}}\ \log \mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big)$

Solution (A5) $\implies$ $\mathrm{y}|\mathrm{X} \sim \text{MVN}(\mathrm{X}\beta, \sigma^2 I_n)$. Therefore $\mathcal{L}(\mathrm{y}|\mathrm{X},\beta,\sigma^2) = (2\pi\sigma^2)^{\frac{-n}{2}} e^{-\frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2\sigma^2}}$, and $\log \mathcal{L}(\mathrm{y}|\mathrm{X},\beta,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{(\mathrm{y}-\mathrm{X}\beta)'(\mathrm{y}-\mathrm{X}\beta)}{2\sigma^2}$. The two FOCs of the maximization problem give an exact closed-form solution:[7]

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

Properties (assuming (A1)-(A5))

- Finite samples

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

  efficient $\mathbb{V}\big[\hat{\beta}_{\text{ML}}|\mathrm{X}\big] = \mathbb{E}\Big[\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)\big(\hat{\beta} - \mathbb{E}[\hat{\beta}]\big)' \mid \mathrm{X}\Big] = \mathbb{E}\Big[\big(\hat{\beta} - \beta\big)\big(\hat{\beta} - \beta\big)' \mid \mathrm{X}\Big]$
  $= \mathbb{E}\big[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{e}\big((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{e}\big)'|\mathrm{X}\big]$
  $= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\ \mathbb{E}[\mathrm{e}\mathrm{e}'|\mathrm{X}]\ \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1} = \sigma^2(\mathrm{X}'\mathrm{X})^{-1} \leqslant \mathbb{V}\big[\hat{\beta}_{...}|\mathrm{X}\big]$

  normally distributed $\hat{\beta}_{\text{ML}} = \beta + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{e} \sim \mathcal{N}\big(\beta, \sigma^2(\mathrm{X}'\mathrm{X})^{-1}\big)$

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

- Asymptotics

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

  asymptotically efficient as is efficient

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

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

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

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

[7] 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.

3 Post-estimation model diagnostics

After our statistical software has fit the model and produced the estimates requested, we should check that the assumptions underlying these numbers hold. Several key assumptions can be diagnosed by simply looking at the residuals. Indeed, these contain the variation in y and the features of the relationship between y and X that weren't explained by the model.

The four plots presented below provide information w.r.t. an assumption of the CLRM.[8] In each figure, the "Case 1" plot illustrates a case where the given assumption seems to be met relatively well, while the "Case 2" plot suggests the reverse.

1. **"Residuals vs Fitted" plot** — *Is there an unmodeled non-linear pattern?*

   Residuals are plotted against fitted values. Are the residuals spread rather equally around a horizontal line, without distinct patterns? If they aren't, it suggests that a non-linear relationship was not explained by the model and was therefore left out in the residuals. *Note: If y_i is discrete, such as in a logistic regression, then residuals are discrete. One shouldn't plot raw residuals, but rather binned residuals (divide the data equally into bins based on fitted values, s.t. each bin has the same number of points, and take the averages for each bin).*

2. **"Scale-Location" plot** — *Are the residuals homoscedastic?*

   The square root of the absolute value of standardized residuals $\sqrt{|r_i|}$ is plotted against fitted values $\hat{y}_i$. Is the vertical spread of points uniform along x? A uniform spread indicates residuals have a uniform variance across the range of predicted values. The reverse suggests there is heteroscedasticity.

3. **"Normal $\mathbb{Q}$-$\mathbb{Q}$" plot** — *Are the residuals normally distributed?*

   The quantiles of the residuals are plotted against the theoretical quantiles of the normal distribution. If the residuals are approximately normally distributed, we should see a roughly straight line. The plot may also reveal outliers.

4. **"Residuals vs Leverage" plot** — *Are there influential observations?*

   First, let's distinguish outliers, high leverage points, and influential points:

   - Outliers are observations with unusual outcome values y_i (i.e., that are considerably different from the rest of the data). They may not have a lot of influence on the regression line.
   - High-leverage points are observations with unusual predictor values x_i. In linear regression, leverage measures how sensitive a fitted $\hat{y}_i$ is to a change in the true y_i. High-leverage points will not have a lot of influence on the regression line if they lie close to it.
   - Finally, influential points are observations whose removal from the data would cause a large change in the estimated regression line. I.e., they largely disagree with the trend.

   A point has to have at least some leverage in order to be influential. To identify influential points, we can compute each observation's Cook's distance d_i, which measures the effect of omitting that observation on the parameter vector. Precisely, d_i has a component that reflects how well the model fits the i-th observation y_i and a component that measures how far that point is from the rest of the data. Points with $d_i > 1$ are generally considered influential.
   In the plot, residuals are plotted against their leverage, and dotted red lines represent Cook's distances of 0.5 and 1. Points outside these lines have high Cook's distances, i.e., high influence.[9]

[8] These 4 particular types of plots are particularly easy to produce: they are built-in diagnostic plots for linear regression analysis in R (one need simply run `plot()` on the fitted model object). The figures used here for illustrative purposes are taken from https://data.library.virginia.edu/diagnostic-plots/.

[9] If the lines aren't visible on the graph, it means that all points are well inside them — there are no influential points.

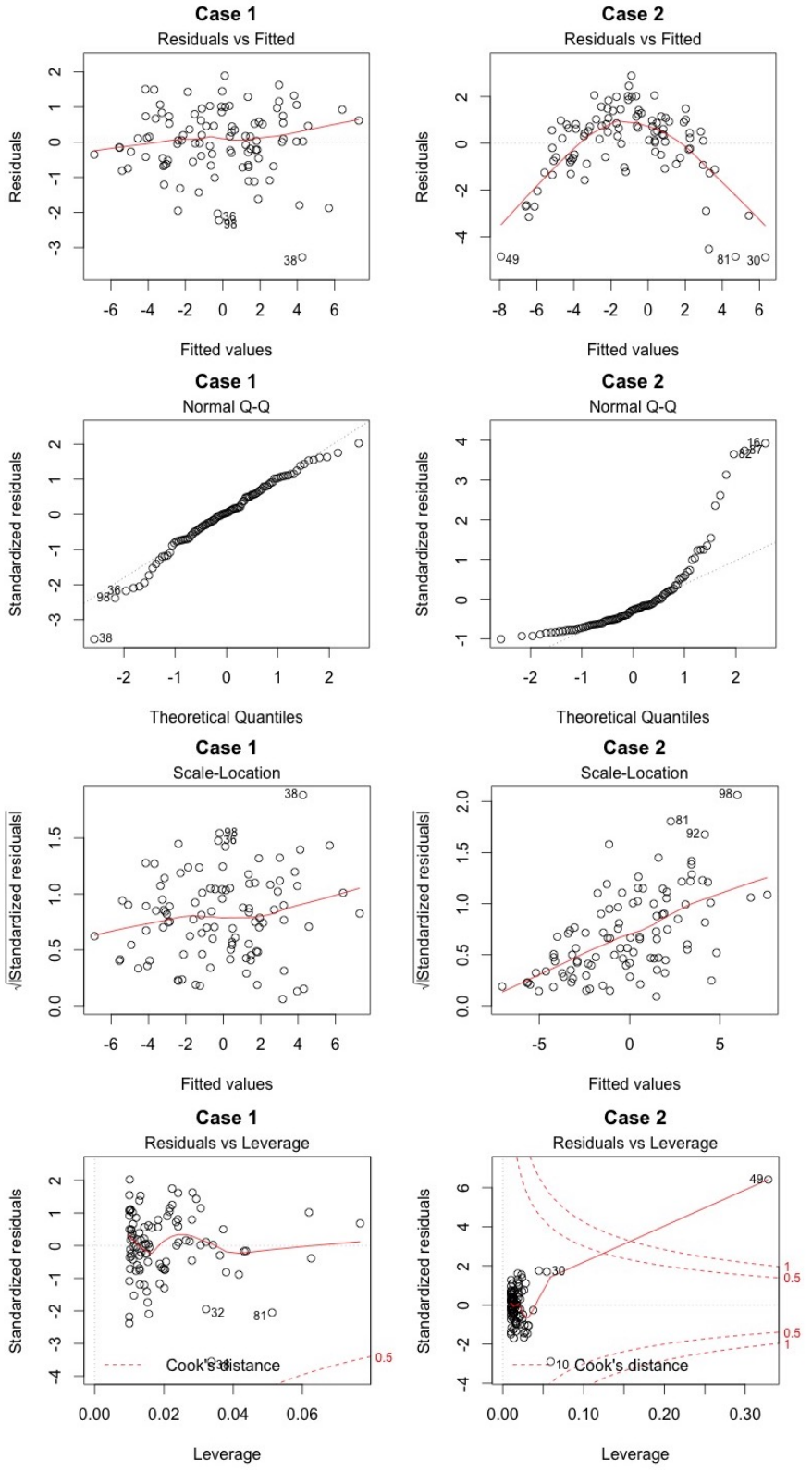


Figure 1: Diagnostic plots. From top to bottom: Residuals vs Fitted; Scale-Location; Normal ℚ-ℚ; Residuals vs Leverage.

4 How to deal with non-spherical errors in OLS

4.1 Sandwich estimators

Assuming (A1)-(A3), by applying the CLT, we obtain the limit distribution of the rescaled $\hat{\beta}_{\text{OLS}}$:[10]

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

where
- $M_{\mathrm{XX}} := \operatorname{plim}\big(\frac{1}{n}\mathrm{X}'\mathrm{X} \mid \mathrm{X}\big) \overset{[11]}{=} \lim\big(\mathbb{E}\big[\frac{1}{n}\mathrm{X}'\mathrm{X}|\mathrm{X}\big]\big) = \lim\big(\frac{1}{n}\mathrm{X}'\mathrm{X}\big)$ is finite and $\neq 0$
- $M_{\mathrm{X}\Sigma\mathrm{X}} := \operatorname{plim}\big(\frac{1}{n}\mathrm{X}'\mathrm{e}\mathrm{e}'\mathrm{X} \mid \mathrm{X}\big) = \lim\big(\mathbb{E}\big[\frac{1}{n}\mathrm{X}'\mathrm{e}\mathrm{e}'\mathrm{X}|\mathrm{X}\big]\big) = \lim\big(\frac{1}{n}\mathrm{X}'\ \mathbb{E}\big[\mathrm{e}\mathrm{e}'|\mathrm{X}\big]\mathrm{X}\big) := \lim\big(\frac{1}{n}\mathrm{X}'\Sigma\mathrm{X}\big)$
- Σ is the variance-covariance matrix of the error term: $\mathbb{E}[\mathrm{e}\mathrm{e}'|\mathrm{X}]$

We talk of the limit distribution of $\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta\big)$, 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 carrying around $\sqrt{n}\big(\hat{\beta}_{\text{OLS}} - \beta\big)$. 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:

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

We need a consistent estimator of the asymptotic variance-covariance matrix $\overset{\text{a}}{\mathbb{V}}[\hat{\boldsymbol{\beta}}_{\text{OLS}}]$ in order to do sampling-based statistical inference.[12] The only unknown is Σ. We hence need a **consistent** estimator of this Σ.

Error structure

- **Spherical (A4)**

  If assumption (A4) is met, i.e., $\Sigma = \sigma^2\mathrm{I}$, then the asymptotic variance-covariance matrix simplifies to $\overset{\text{a}}{\mathbb{V}} = \big(\mathrm{X}'\mathrm{X}\big)^{-1}\mathrm{X}'\Sigma\mathrm{X}\big(\mathrm{X}'\mathrm{X}\big)^{-1'} = \sigma^2\big(\mathrm{X}'\mathrm{X}\big)^{-1}\big(\mathrm{X}'\mathrm{X}\big)\big(\mathrm{X}'\mathrm{X}\big)^{-1'} = \sigma^2\big(\mathrm{X}'\mathrm{X}\big)^{-1}$. The population variance σ^2 can be consistently estimated by the unbiased sample variance $s^2 = \frac{\sum_i r_i^2}{n-p}$, and hence $\overset{\text{a}}{\mathbb{V}}$ by $\widehat{\overset{\text{a}}{\mathbb{V}}}_{\text{S}} := s^2\big(\mathrm{X}'\mathrm{X}\big)^{-1}$. This expression is actually the Cramer-Rao lower bound, therefore $\hat{\beta}_{\text{OLS}}$ is **BLUE.**

- **Not spherical**

  If assumption (A4) is violated—due to heteroscedasticity or dependence—we need a covariance matrix estimator that is consistent under this misspecification of the remaining likelihood. One approach is

[10] As in the entire document: (i) $\hat{\beta}_{\text{OLS}}$ refers to the vector of both the intercept and the slope coefficients, i.e., $p \geq 2$; (ii) for convenience, we drop the notation |X, however all features of $\hat{\beta}_{\text{OLS}}$'s distribution presented here are actually conditional on X.

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

[12] The standard error used in the t-test for $\hat{\beta}_{\text{OLS}}$ is indeed an estimate of $\sqrt{\overset{\text{a}}{\mathbb{V}}[\hat{\beta}_{\text{OLS}}]}$.

to use **sandwich estimators**.[13] We decompose the variance into its 3 $k \times k$ components: *bread, meat, bread*, and compute a **consistent** estimator of the *meat* component $X'\Sigma X$ that best represents our assumed error structure, to finally compute:

$$\hat{\overset{a}{\mathbb{V}}}\left[\hat{\beta}_{OLS}\right] := \underbrace{\left(X'X\right)^{-1}}_{bread} \underbrace{X'\hat{\Sigma}X}_{meat} \underbrace{\left(X'X\right)^{-1'}}_{bread}$$

- **Heteroskedastic**

  $\Sigma = diag[\sigma_i^2]$. White (1980) proposes to use $\hat{\Sigma}_H := \frac{1}{n-p} diag[r_i^2]$, i.e., $X'\hat{\Sigma}_H X = \frac{1}{n-p}\sum_i r_i^2 x_i x_i'$. The resulting non-parametric estimator $\hat{\overset{a}{\mathbb{V}}}_H$ is consistent for $\overset{a}{\mathbb{V}}$, even though r_i^2 is inconsistent for σ_i^2.

  The resulting standard errors are called heteroskedasticity-consistent (HC) aka "robust". They are 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.[14][15]

- **Autocorrelated**

  If errors are autocorrelated in any way (in time, in space, by groups...), it means that the model is not capturing some feature of the DGP. To conduct proper inference, one can either treat this structure as *substance* and incorporate it in the model (e.g., if errors are autocorrelated by group, by modeling a multilevel data structure), or treat it as *nuisance* and adjust for it after fitting the model (e.g., if errors are autocorrelated by group, by clustering standard errors).[16]

  In the second approach, the dependence structure is left in the errors, such that the *true* covariance matrix of the errors Σ has some non-diagonal terms that are non-zero. We need an estimator $\hat{\Sigma}$ that is consistent for all diagonal and non-diagonal terms. The sandwich estimators below achieve this with the same approach:

  * They assume the process is 2nd order stationary;[17]

[13] All extremum estimators can actually be shown to be consistent and asymptotic normal, with an asymptotic variance matrix in the sandwich form: $\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, A(\theta)^{-1}B(\theta)A(\theta)^{-1}\right)$. The sandwich algorithm presented here for OLS can be extended to all extremum estimators, e.g., ML and GMM. This is not to say that it should be; Freedman (2006) points out that while White's sandwich estimator often gives good results in OLS, the equivalent correction in ML 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 ML 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?)

[14] With a nonlinear conditional expectation function (CEF), the use of a linear model to approximate it should lead to heteroskedasticity (Angrist and Pischke, 2008, p.35). Indeed, as the quality of fit between the regression line and the CEF will vary with x_i, the residuals will be larger, on average, at values of x_i where the fit is poorer. The residual variance will increase with the square of the gap between the regression line $x_i\beta$ and the CEF $\mathbb{E}[y_i|x_i]$.

[15] 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.

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

[17] 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.

* They use sample equivalents: White's estimates for diagonal terms,[18] and sample autocovariances for non-zero non-diagonal terms. They also give weights to these non-zero non-diagonal covariance terms (e.g., through a kernel function).

They are thereby fully non-parametric, and account for dependence of unknown form along the dimension of autocorrelation.

* **Clustering (dimension of autocorrelation: group appartenance)**

  Errors are correlated within groups or "clusters": $\mathbb{E}[e_i|\mathrm{x}_i] = 0$, $\mathbb{E}[e_i e_j|\mathrm{x}_i, \mathrm{x}_j] \neq 0$, $\forall i, j \in$ same group g. The covariance matrix of the error term Σ has a block-diagonal structure.

  The *cluster-robust* estimator uses $\mathrm{X}'\widehat{\Sigma}_{\mathrm{C}}\mathrm{X} = \frac{1}{n-p}\sum_{g=1}^{G}\mathrm{X}_g'\mathrm{r}_g\mathrm{r}_g'\mathrm{X}_g$.

  Note: This estimator is valid only if the number of clusters G is sufficiently large (rule of thumb: > 30), as, like all sandwich estimators, it relies on asymptotics.[19] *Note that if clusters are unbalanced, the effective number of clusters is even lower. Cameron and Miller (2015) recommends using critical values from the t_{G-1} distribution instead of the normal $\mathcal{N}(0,1)$.*

* **Serial correlation (dimension of autocorrelation: time)**

  Newey and West (1987) proposes an estimator that accounts for serial correlation of unknown form in an error time series $\{e_t\}$. It can be expanded to panel datasets by estimating only correlations between lagged errors in the same cluster.

  The time series' autocovariance of lag l is $\gamma_e(t, t-l) := cov[e_t, e_{t-l}] = \mathbb{E}[(e_t - \mu_{e_t})(e_{t-l} - \mu_{e_t})]$. If the process is covariance-stationary, it is a function of the relative lag only: $\gamma_e(l)$. Its bias-adjusted sample equivalent, i.e., the sample autocovariance, is $g(l) = \frac{1}{T-k}\sum_{t=l+1}^{T} r_t r_{t-l}$.

  The Newey-West estimator weights these sample autocovariances using a triangular kernel function,[20] s.t. the weight decreases linearly with the lag up to a chosen maximum lag L, and adds White's variance estimates:

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

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

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

  Note: $\widehat{\Sigma}$ 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 $T^{1/4}$.

[18] These estimators are hence also heteroskedasticity-consistent.

[19] The t-statistic $t_{\hat{\beta}} = \dfrac{\hat{\beta} - \beta}{\sqrt{\widehat{\mathbb{V}}_{\mathrm{C}}[\hat{\beta}]}} \underset{H_0}{\overset{a}{\sim}} \mathcal{N}(0,1)$. However, for finite G (and therefore, especially for small $G < 30$), $t_{\hat{\beta}}$'s distribution is unknown — even with normal errors. Intuitively, fewer clusters means there is less independent information in the sample (as the data are independent across clusters but not within). 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.

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

∗ **Spatial correlation (dimension of autocorrelation: 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. It weights the sample covariances using a kernel function $k(s_i, s_j)$, where s_i is the location of observation i.

$$\mathrm{X}'\widehat{\Sigma}_{\text{Co}}\mathrm{X} \coloneqq \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 \mathrm{x}_i\mathrm{x}_i' + \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j=1}^{n} k(s_i, s_j)\; r_i r_j \mathrm{x}_i\mathrm{x}_j'$$

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, $k(i, j)$ simplifies to a function of distance: $k(d_{ij})$. One can choose whichever distance metric fits one's context, e.g., a metric of economic distance.*
- *Conley (1999) shows that spatial dependence does not imply that SEs will necessarily increase. In his empirical example, 6 out of 9 spatial SE estimates are smaller than their iid counterparts.*
- *Similarly to cluster robust standard errors, these perform well only when there is a reasonable effective number of independent clusters, which decreases as the radius is extended.*
- *This estimator is very similar to the method of Kriging in geostatistics.*

⚠ **Sandwich estimators are pointless in ML estimation** These computations of adjusted SEs make sense only for the *linear* regression model estimated by OLS, where the OLS point estimator remains unbiased (but is not "best" in the sense of having minimum mean square error), and they serve to address that the OLS variance estimator would not be consistent for the variance of the OLS estimates. In the case of a model that is nonlinear in the parameters (e.g., a Logit or Probit model, which is usually estimated by ML), if for example the errors are heteroskedastic, then:

- The ML estimator of $\overset{\text{a}}{\mathbb{V}}[\hat{\beta}_{\text{ML}}]$ is inconsistent (as in the linear model);
- But $\hat{\beta}_{\text{ML}}$ itself is also biased (in an unknown direction), and inconsistent (unless the likelihood function is modified to correctly take into account the precise form of heteroskedasticity).

The reporting of robust standard errors in the context of nonlinear models such as Logit and Probit therefore doesn't make sense. What use is a consistent SE when the point estimate is wrong (Freedman, 2006)? This is why it is important to test for model mis-specification (such as heteroskedasticity) when estimating models such as Logit, Probit, Tobit...[21] Then, if need be, the model can be modified to take the heteroskedasticity into account before we estimate the parameters; e.g., using fixed or random effects.

4.2 Block bootstrap

- Bootstrapping is a resampling technique which allows estimation of the sampling distribution of almost any statistic.
- Goal: estimating the sampling distribution of a statistic (ex of statistic: a regression coefficient)

[21] The reason why one can use a sandwich estimator in a linear model is because the coefficients and standard errors are determined separately. In nonlinear models estimated by ML, the coefficients and standard errors can't be separated, they are jointly determined by maximizing the likelihood of y|X. The problem applies to most of the standard models (binary, multinomial, ordered, and count data models) with the exception of GLS and Poisson.

- Assumption: Bootstrap relies on the assumption that the data represent the actual underlying (population) distribution well. It doesn't make specific distributional assumptions.
- Method: create an approximating distribution: from the existing original sample of size n, take a random sample (with replacement) also of size n. Do this B times. From each of the B boostrap resamples, compute the statistic of interest $\hat{\beta}_b$. We obtain a distribution $(\hat{\beta}_1, ..., \hat{\beta}_B)$.
- Clustered Data
  As observations within the same cluster tend to be more alike with each other compared with observations in other clusters (e.g., students in the same class have a common teacher), observations exhibit some degree of interdependence. This interdependence is a result of the sampling design typically found in CRTs where all students in one group or cluster are assigned to a condition which then affects the variance of the outcome which in turn affects the estimates of the standard errors. While OLS point estimates should be unbiased, the greater concern when dealing with clustered data revolves around the standard errors.
  Cluster bootstrapping: (Good alternative esp. when the number of clusters is small.) Our data has n observations across J clusters. Each of the B resamples is obtained by: randomly selecting J clusters with replacement (so some clusters will be selected more than once and others not selected at all), and including all observations within these clusters in the overall bootstrapped sample. If clusters have different sizes, the bootstrapped sample b may not be of size n. The block bootstrap is used when the data, or the errors in a model, are correlated. In this case, a simple case or residual resampling will fail, as it is not able to replicate the correlation in the data. The block bootstrap tries to replicate the correlation by resampling inside blocks of data. The block bootstrap has been used mainly with data correlated in time (i.e. time series) but can also be used with data correlated in space, or among groups (cluster data). Cluster data describes data where many observations per unit are observed. This could be observing many firms in many states or observing students in many classes. In such cases, the correlation structure is simplified, and one does usually make the assumption that data is correlated within a group/cluster, but independent between groups/clusters. The structure of the block bootstrap is easily obtained (where the block just corresponds to the group), and usually only the groups are resampled, while the observations within the groups are left unchanged.

References

Angrist, J. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, Princeton, NJ, ISBN: 9781400829828, DOI: 10.1515/9781400829828.

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

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

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

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

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

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

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

A Misc.

A.1 Deriving the formula of the OLS estimator

Consider the multivariate linear regression model. We can write it as a system of n equations, or equivalently, in its matrix form:

$$y_i = \mathrm{x}_i'\beta + e_i = \sum_{j=0}^{k} \beta_j x_{ij} + e_i, \quad e_i \overset{\text{iid}}{\sim} (0, \sigma^2), \quad \text{for } i = 1, ..., n$$

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{e}, \quad \mathrm{e} \sim (0, \sigma^2 \mathrm{I}_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The OLS estimator is defined as the minimizer of the sum of squared residuals: $\hat{\beta}_{\text{OLS}} \coloneqq \underset{\beta}{\operatorname{argmin}}\ \text{SSR} \coloneqq \underset{\beta}{\operatorname{argmin}}\ \sum_{i=1}^{n} r_i^2 = \underset{\beta}{\operatorname{argmin}}\ \sum_{i=1}^{n} (y_i - \mathrm{x}_i'\beta)^2$. We can solve for $\hat{\beta}_{\text{OLS}}$ using calculus:

- **Matrix form**

$$\hat{\beta}_{\text{OLS}} \coloneqq \underset{\beta}{\operatorname{argmin}}\ (\mathrm{y} - \mathrm{X}\beta)'(\mathrm{y} - \mathrm{X}\beta) = \underset{\beta}{\operatorname{argmin}}\ \left(\mathrm{y}'\mathrm{y} - \beta'\mathrm{X}'\mathrm{y} - \mathrm{y}'\mathrm{X}\beta + \beta'\mathrm{X}'\mathrm{X}\beta\right)$$

$$\begin{aligned} \text{FOC:} \quad \frac{dS}{d\beta}(\hat{\beta}) = 0 &\iff \frac{d}{d\beta}\left(\mathrm{y}'\mathrm{y} - \beta'\mathrm{X}'\mathrm{y} - \mathrm{y}'\mathrm{X}\beta + \beta'\mathrm{X}'\mathrm{X}\beta\right)\Big|_{\beta=\hat{\beta}} = 0 \\ &\iff -\mathrm{X}'\mathrm{y} - (\mathrm{y}'\mathrm{X})' + 2\mathrm{X}'\mathrm{X}\beta\big|_{\beta=\hat{\beta}} = 0^{22} \\ &\iff -2\mathrm{X}'\mathrm{y} + 2\mathrm{X}'\mathrm{X}\hat{\beta} = 0 \\ &\iff \hat{\beta} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y} \end{aligned}$$

- **System of n equations**

$$\hat{\beta}_{\text{OLS}} \coloneqq \underset{\beta}{\operatorname{argmin}} \sum_i r_i^2 = \underset{\beta}{\operatorname{argmin}} \sum_i \left(y_i - \sum_k \beta_k x_{ik}\right)^2$$

$$\begin{aligned} \text{FOC:} \quad \forall j, \frac{\partial \sum_i r_i^2}{\partial \beta_j} = 0 &\iff 2\sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0 \\ &\iff \sum_i \left(y_i - \textstyle\sum_k \beta_k x_{ik}\right) \dfrac{\partial\left(y_i - \sum_k \beta_k x_{ik}\right)}{\partial \beta_j} = 0 \\ &\iff \sum_i \left(y_i - \textstyle\sum_k \beta_k x_{ik}\right)(-x_{ij}) = 0 \\ &\iff \sum_i x_{ij} y_i = \sum_i x_{ij} \textstyle\sum_k \beta_k x_{ik} \end{aligned}$$

For example:

[22] Using denominator-layout notation, we have the following derivatives, or scalar-by-vector identities (where β and A are vectors): $\frac{d\beta' A}{d\beta} = \frac{dA'\beta}{d\beta} = A$, $\frac{d\beta' A\beta}{d\beta} = 2A\beta$.

– For the univariate regression model $y_i = \beta_0 + \beta_1 x_i + e_i$ (i.e., $x_{i0} = 1$ and $x_{i1} = x_i$):
  * $\hat{\beta}_0$: $\sum_i y_i = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \iff n\bar{y} = n\hat{\beta}_0 + n\hat{\beta}_1 \bar{x} \iff \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  * $\hat{\beta}_1$: $\sum_i x_i y_i = \sum_i x_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \iff \sum_i x_i y_i = \hat{\beta}_0 n\bar{x} + \hat{\beta}_1 \sum_i x_i^2$
  $\iff \frac{1}{n}\sum_i x_i y_i = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right)\bar{x} + \hat{\beta}_1 \frac{1}{n}\sum_i x_i^2$
  $\iff \frac{1}{n}\sum_i (x_i y_i) - \bar{y}\bar{x} = \hat{\beta}_1 \left(\frac{1}{n}\sum_i (x_i^2) - \bar{x}^2\right)$
  $\iff \hat{\beta}_1 = \frac{\frac{1}{n}\sum_i (x_i y_i) - \bar{y}\bar{x}}{\frac{1}{n}\sum_i (x_i^2) - \bar{x}^2} = \ldots = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$

  This is the sample equivalent of the estimand: $\beta_1 = \frac{\text{cov}[x,y]}{\mathbb{V}[x]}$.
– For the bivariate regression model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$
  * $\hat{\beta}_0$: ... $\iff \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_2 \bar{z}$
  * $\hat{\beta}_1$: ... $\iff \hat{\beta}_1$ = the sample analog of $\frac{\text{cov}[x,y]\mathbb{V}[z] - \text{cov}[x,z]\text{cov}[z,y]}{\mathbb{V}[x]\mathbb{V}[z] - \text{cov}[x,z]^2}$
  * $\hat{\beta}_2$: ... $\iff \hat{\beta}_2$ = the sample analog of $\frac{\text{cov}[z,y]\mathbb{V}[x] - \text{cov}[x,z]\text{cov}[x,y]}{\mathbb{V}[x]\mathbb{V}[z] - \text{cov}[x,z]^2}$
– For the multivariate regression model $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i$
  * $\hat{\beta}_k$: ... $\iff \hat{\beta}_k$ = the sample analog of $\frac{\text{cov}[\tilde{x}_k, y]}{\mathbb{V}[\tilde{x}_k]}$ where $\tilde{x}_k$ is the residual from the regression of x_k on all the other covariates.

A.2 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.3 Kernel functions for non-parametric statistics

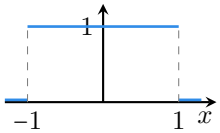
A **kernel** is a non-negative real-valued integrable function $k()$, used as a **weighting function** in non-parametric estimation techniques. They are also called "window functions" (notably in time-series).

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

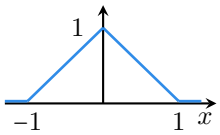
- normalization: $\int_{-\infty}^{+\infty} k(u)\,du = 1$
  In kernel density estimation, this ensures that the estimation produces a probability density function.
- symmetry: $\forall u,\ k(-u) = k(u)$
  This ensures that the average of the corresponding distribution is equal to that of the sample used.

Examples of commonly used symmetric kernels, with the arbitrary bounded support $[-1, 1]$:

- uniform: $k(u) = \begin{cases} 1 & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$



- triangular (Bartlett): $k(u) = \begin{cases} 1 - |u| & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$



- modified Bartlett (*a la Newey-West*): $k(u) = \begin{cases} 1 - \dfrac{u}{L+1} & \text{if } |u| \leq L \\ 0 & \text{otherwise} \end{cases}$
- parabolic (Epanechnikov): $k(u) = \begin{cases} \dfrac{3}{4}\left(1 - u^2\right) & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$