Lecture 16: Elements of Likelihood Theory

Econometrics 2 – *Parametric Models, KL Divergence, and the MLE*

Instructor: Prof. S. Arvanitis | **Notes & digitisation:** T. Kourtalis

Semester: Spring 2026

▷ Amber = Handwritten Notes (professor's words) ◇ Teal = Student's Notes

Recall from Earlier Lectures

Where we are in the course:

- **Lectures 2–5:** Semi-parametric models; OLS as an M-estimator; consistency and asymptotic normality framework.
- **Lectures 6–15:** IV and GMM estimators; consistency and normality of OLS and IV completed.

This lecture introduces fully parametric models, where the entire distribution of the sample is specified up to a finite-dimensional parameter. This allows us to define the likelihood, connect MLE to KL divergence minimisation, and show that MLE equals OLS under Gaussian errors.

1 Elements of Likelihood Theory

▷ **Handwritten Notes** (what the professor said)

Likelihood theory applies to parametric models. To each $\theta \in \Theta$ corresponds a unique distribution. Let P_θ be the distribution for the sample. This can be represented by:

- a probability mass function (when it is discrete).
- a probability density function (when it is continuous or smooth).

For the following, we assume that the statistical model is correctly specified, meaning $\exists\, \theta_0 \in \Theta$ and P_{θ_0} is the true distribution of the sample.
In a parametric model, for each value of the parameter, β is a well-defined function. Let:

$$\begin{aligned} p(\beta, z_n) = p(\beta, z_{(n)} \mid z_{(n-1)}, \ldots, z_{(1)}) \\ \cdot\, p(\beta, z_{(n-1)} \mid z_{(n-2)}, \ldots, z_{(1)}) \cdots p(\beta, z_{(1)}) \end{aligned}$$

In the case of independence, the factors are the marginal distributions of $z_{(i)}$ with respect to the parameter β.

◇ Student's Notes

This lecture marks a clean break from Lectures 2–15. Everything up to now was semi-parametric: we specified moments ($\mathbb{E}(\varepsilon \mid X) = 0$, $\mathrm{Var}(\varepsilon \mid X) = I$) but left the full shape of the distribution of ε unspecified. Now we commit to specifying everything. The payoff is the likelihood function and the MLE, which under regularity conditions achieves the Cramer-Rao lower bound and is asymptotically efficient.
The chain rule decomposition is worth staring at. It is just $P(\text{all data}) = P(\text{last obs} \mid \text{all previous}) \times P(\text{second to last} \mid \text{all earlier}) \times \cdots$. Under i.i.d. this collapses to $\prod_{i=1}^{n} p(\beta, z_{(i)})$ and order does not matter. Under dependence (time series, panel) the conditioning structure is real and ordering matters.

Feature	Semi-parametric (Lects 2–15)	Parametric (this lecture)
What is specified	Moments only	Full distribution P_θ
Objective function	$M_n(\beta)$: OLS, IV, LAD	Log-likelihood $\ell_n(\theta)$
Estimator	M-estimator	MLE
Efficiency	Not necessarily optimal	Cramer-Rao efficient
Misspecification cost	Low: moments may still hold	High: wrong family ruins everything

2 The Linear Model as a Parametric Model

▷ Handwritten Notes (what the professor said)

Let the linear model be:

$$Y_n = X_n\beta + \varepsilon_n, \qquad \varepsilon_n \mid X_n \sim N(0_{n\times 1}, I_{n\times n})$$

Therefore $Y_n \mid X_n \sim N(X_n\beta, I_{n\times n})$. For each observation i: $Y_{(i)} \mid X_n \sim N(X_{(i)}\beta, 1)$, independent of $Y_{(j)}$ for $i \neq j$ given X_n. We set $Z_n = Y_n$, $z_{(i)} = Y_{(i)}$.
The joint distribution is:

$$p(\beta, z_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\tfrac{1}{2}(Y_n - X_n\beta)'(Y_n - X_n\beta)\right)$$

Due to conditional independence:

$$p(\beta, z_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}(Y_{(i)} - X_{(i)}\beta)^2\right)$$

◇ Student's Notes

The assumption $\varepsilon_n \mid X_n \sim N(0, I)$ is what turns the linear model from semi-parametric into fully parametric. In Lectures 2–14 we only needed $\mathbb{E}(\varepsilon \mid X) = 0$ for consistency and $\mathbb{E}(\varepsilon^2 X'X) < \infty$ for normality. Now we commit to the full Gaussian shape. This is a stronger assumption and it buys us efficiency.

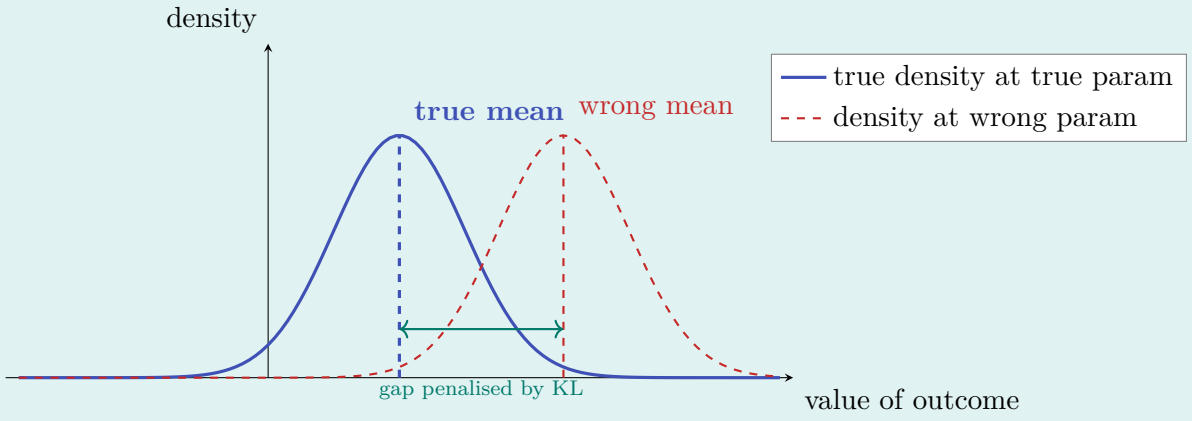

Figure 1: The conditional density of $Y_{(i)}$ at the true parameter (indigo) versus a wrong parameter (rose dashed). The KL divergence measures how far the wrong density is from the true one. Maximising the likelihood is equivalent to minimising this gap.

3 Kullback-Leibler Divergence

▷ Handwritten Notes (what the professor said)

The assumption of correct specification is equivalent to $p(\beta_0, z_n)$ being the true distribution. Expected values with respect to this are denoted $\mathbb{E}_{\beta_0}$.
We have:

$$KL(p(\beta_0, z_n),\, p(\beta, z_n)) := \mathbb{E}_{\beta_0}(\log p(\beta_0, z_n)) - \mathbb{E}_{\beta_0}(\log p(\beta, z_n))$$

Analysing the second term:

$$\mathbb{E}_{\beta_0}(\log p(\beta, z_n)) = \mathbb{E}_{\beta_0}\left(\sum_{i=1}^{n} \log p(\beta, z_{(i)} \mid z_{(i-1)}, \ldots, z_{(1)})\right)$$

The above quantity is uniquely maximised at β_0. It is not observable because it involves $\mathbb{E}_{\beta_0}$, which depends on the unknown true distribution. We approximate it by:

$$\frac{1}{n}\sum_{i=1}^{n} \log p(\beta, z_{(i)} \mid z_{(i-1)}, \ldots, z_{(1)})$$

◇ Student's Notes

The KL divergence is one of the most important objects in statistics and information theory. It deserves a careful treatment.

Why KL is always non-negative.
The proof uses Jensen's inequality. Since log is strictly concave, for any random variable $R > 0$:

$$\mathbb{E}(\log R) \leq \log \mathbb{E}(R)$$

Apply this with $R = p(\beta, z_n)/p(\beta_0, z_n)$:

$$\mathbb{E}_{\beta_0}\left(\log \frac{p(\beta, z_n)}{p(\beta_0, z_n)}\right) \leq \log \mathbb{E}_{\beta_0}\left(\frac{p(\beta, z_n)}{p(\beta_0, z_n)}\right) = \log \int p(\beta, z_n)\, dz_n = \log 1 = 0$$

Rearranging:

$$\mathbb{E}_{\beta_0}(\log p(\beta_0, z_n)) - \mathbb{E}_{\beta_0}(\log p(\beta, z_n)) \geq 0 \implies KL \geq 0$$

with equality if and only if $p(\beta, \cdot) = p(\beta_0, \cdot)$ almost everywhere, which under identifiability means $\beta = \beta_0$. This is why maximising $\mathbb{E}_{\beta_0}(\log p(\beta, z_n))$ uniquely identifies β_0.

KL is not a distance.
Despite being called a divergence and measuring how far $p(\beta, \cdot)$ is from $p(\beta_0, \cdot)$, KL is not a metric. It fails two of the three metric axioms:

- **Not symmetric:** $KL(p, q) \neq KL(q, p)$ in general. $KL(p(\beta_0, \cdot), p(\beta, \cdot))$ measures the cost of approximating $p(\beta_0, \cdot)$ by $p(\beta, \cdot)$, which is different from the cost in the reverse direction.
- **No triangle inequality:** $KL(p, r) \leq KL(p, q) + KL(q, r)$ need not hold.

It does satisfy $KL(p, q) \geq 0$ with equality iff $p = q$. For MLE purposes the asymmetry is exactly right: we want to find β that minimises the cost of replacing the true $p(\beta_0, \cdot)$ with the model $p(\beta, \cdot)$, not the other way around.

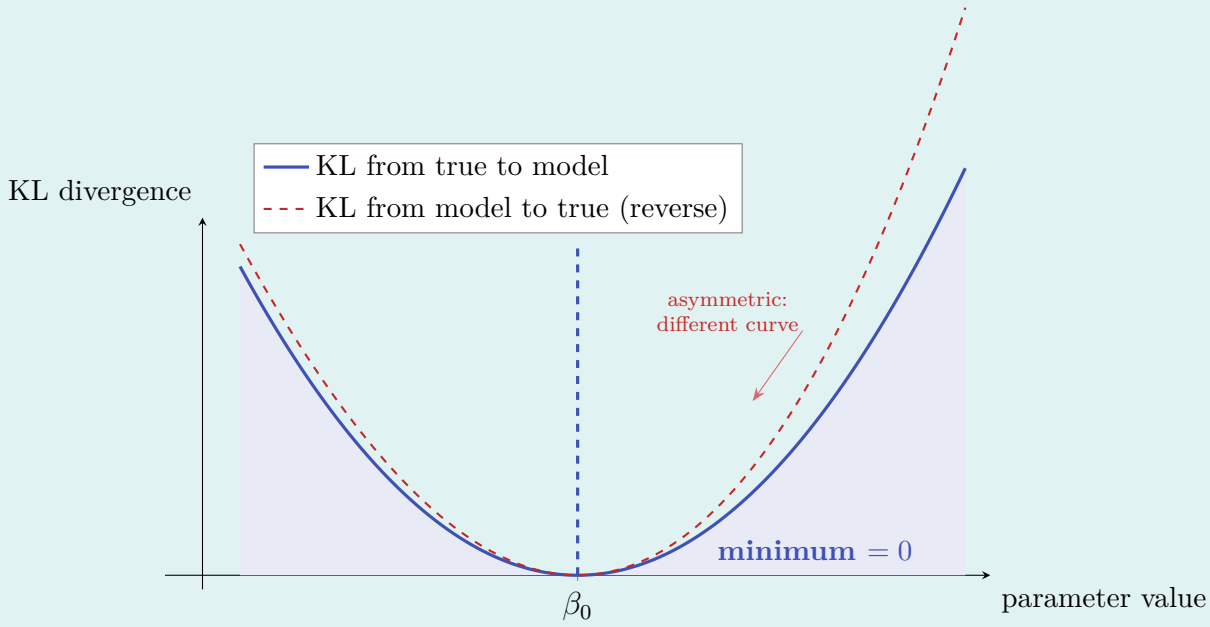

Figure 2: The KL divergence from the true distribution to the model (indigo, filled) is always non-negative and hits zero only at β_0. The reverse KL (rose dashed) has the same zero but a different shape, illustrating the asymmetry of KL. MLE minimises the forward KL (indigo), not the reverse.

▷ Edge Cases and Pathologies

Edge Case 1: KL is infinite when supports do not overlap.
If the true distribution $p(\beta_0, \cdot)$ assigns positive probability to a region where the model distribution $p(\beta, \cdot)$ assigns zero probability, then $\log p(\beta, z_n) = -\infty$ on that region, and:

$$KL(p(\beta_0, \cdot), p(\beta, \cdot)) = +\infty$$

This is a genuine disaster: the model cannot even represent events that can actually happen. In the MLE context this means the log-likelihood is $-\infty$ at such β, so those values are automatically excluded from the argmax. But it also means the analogy principle and the LLN argument for consistency break down, because $\log p(\beta, z_n)$ may not have a finite expectation.
A concrete example: if the true errors are t-distributed with 3 degrees of freedom (heavy tails) but the model assumes Gaussian errors, then both distributions have full support on $\mathbb{R}$ and KL is finite. But if the model assumes uniform errors on $[0, 1]$ and the truth has any probability outside that interval, KL is infinite and MLE fails.

Edge Case 2: KL equals zero at multiple points (non-identification).
If $p(\beta_1, \cdot) = p(\beta_0, \cdot)$ for some $\beta_1 \neq \beta_0$, then $KL = 0$ at both β_0 and β_1. The model is **not identified**: different parameter values produce the same distribution, so data cannot

distinguish between them. MLE then has multiple maximisers and is inconsistent in the sense that it does not converge to a unique limit.
This connects directly to the identification condition we have used throughout the course: $\text{rank}(Q_{X'X}) = p$ in OLS (Lecture 11), $\text{rank}(Q_{Z'X}) = p$ in IV (Lecture 12). Both conditions prevented the limit objective function from having flat regions, ensuring unique minimisers. The same logic operates here through KL.

Edge Case 3: Misspecified models and the pseudo-true value.
What if the model is misspecified, i.e., no $\beta \in \Theta$ satisfies $p(\beta, \cdot) = p(\beta_0, \cdot)$? Then $KL(p(\beta_0, \cdot), p(\beta, \cdot)) > 0$ for all $\beta \in \Theta$. But KL still has a minimum over Θ, call it β^*:

$$\beta^* = \arg\min_{\beta \in \Theta} KL(p(\beta_0, \cdot), p(\beta, \cdot)) = \arg\max_{\beta \in \Theta} \mathbb{E}_{\beta_0}(\log p(\beta, z_n))$$

This β^* is the **pseudo-true value** under misspecification. The MLE will converge to β^*, not to the true β_0. This is the exact same concept as the pseudo-true value under mild misspecification from Lecture 11 for OLS (the projection of β_0 onto Θ under the $Q_{X'X}$ metric). The KL framework gives a general theory of what estimators converge to when the model is wrong.

◇ Student's Notes

Why the approximation step is valid.
The professor replaces the unobservable $\mathbb{E}_{\beta_0}(\log p(\beta, z_n))$ with the sample average $\frac{1}{n}\sum_{i=1}^{n} \log p(\beta, z_{(i)} \mid z_{(i-1)}, \ldots, z_{(1)})$. This is the analogy principle from Lecture 4, applied to the log-likelihood. The LLN (Kolmogorov, under i.i.d. and existence of $\mathbb{E}(|\log p(\beta, z_{(i)})|)$) guarantees:

$$\frac{1}{n}\sum_{i=1}^{n} \log p(\beta, z_{(i)}) \xrightarrow{p} \mathbb{E}_{\beta_0}(\log p(\beta, z_{(i)}))$$

for each fixed β. Local uniform convergence (needed for consistency) requires additional regularity on $\log p$ as a function of β, which is where the smoothness conditions for MLE come from.

KL and information.
The KL divergence has a beautiful interpretation in information theory. $KL(p(\beta_0, \cdot), p(\beta, \cdot))$ is the expected number of extra bits needed to encode data drawn from $p(\beta_0, \cdot)$ using a code optimised for $p(\beta, \cdot)$. Minimising KL means finding the distribution in the model family that is closest to the truth in the information-theoretic sense. This is why MLE is the natural estimation method for parametric models: it finds the member of the parametric family that best compresses the observed data.

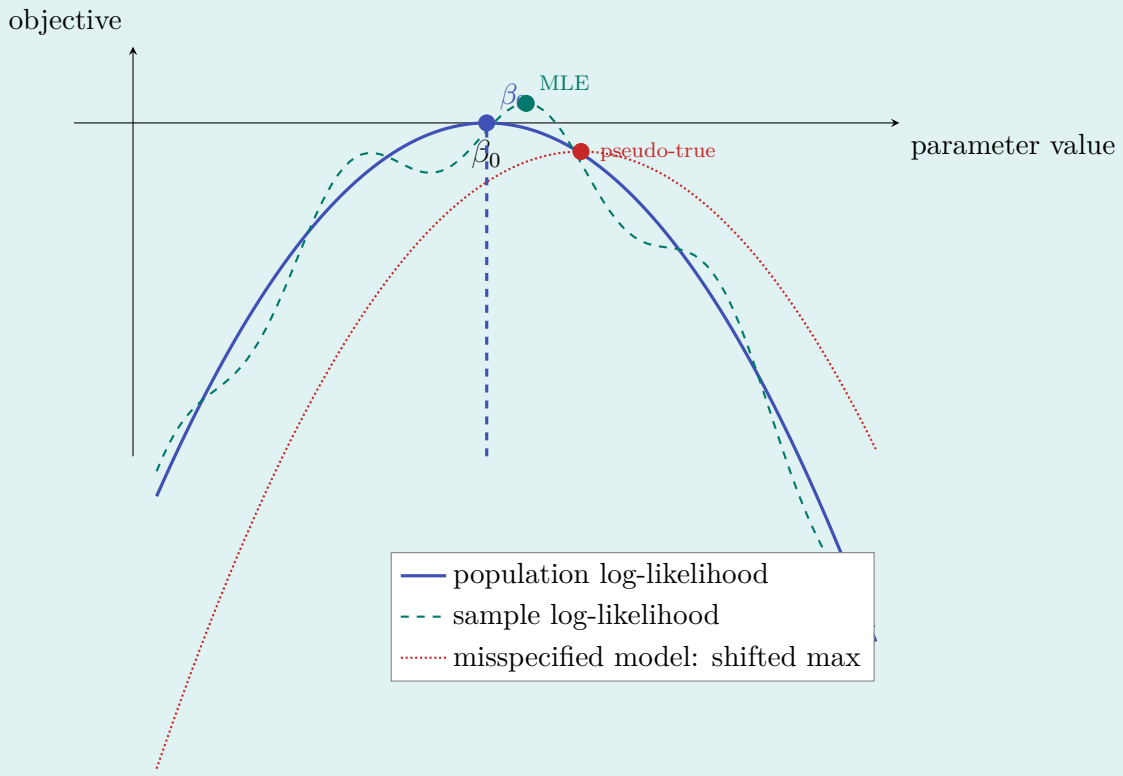

Figure 3: Three scenarios. The population log-likelihood (indigo) is maximised at β_0. The sample version (teal dashed) gives an MLE close to but not at β_0 for finite n. The misspecified model (rose dotted) has its maximum at $\beta^* \neq \beta_0$, the pseudo-true value: the model parameter that is closest to the truth in the KL sense. The MLE under misspecification converges to β^*, not β_0.

4 Maximum Likelihood Estimator

▷ Handwritten Notes (what the professor said)

The Maximum Likelihood Estimator of β_0 is:

$$\hat{\beta}_n = \arg\max_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p(\beta, z_{(i)} \mid z_{(i-1)}, \ldots, z_{(1)})$$

For the linear model:

$$\frac{1}{n} \log p(\beta, z_n) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2n} \sum_{i=1}^{n} (Y_{(i)} - X_{(i)}\beta)^2$$

◇ Student's Notes

The MLE is just an M-estimator with objective function equal to the average log-likelihood. All the general theory from Lectures 10–13 applies directly: the population criterion $\mathbb{E}_{\beta_0}(\log p(\beta, z_n))$ is uniquely maximised at β_0 by the KL argument, the sample average approximates it by the LLN, and the two-step consistency recipe from Lecture 10 gives consistency. Asymptotic normality follows from the Taylor expansion framework of Lecture 13 under smoothness of $\log p$.
The score function (gradient of the log-likelihood) plays the role of $\frac{\partial \mu_n(\beta_0)}{\partial \beta}$ in the general framework. Under regularity conditions it has mean zero at β_0 (because $\frac{d}{d\beta}\int p(\beta, z)\, dz = 0$ implies $\mathbb{E}_{\beta_0}(\nabla_\beta \log p(\beta_0, z_n)) = 0$). The Fisher information matrix $\mathcal{I}(\beta_0) = \mathbb{E}_{\beta_0}((\nabla_\beta \log p(\beta_0, z))(\nabla_\beta \log p(\beta_0, z))')$ plays the role of V in the sandwich. And for the MLE, a special cancellation occurs: $H(\beta_0) = -\mathcal{I}(\beta_0)$, so the sandwich $H^{-1}VH^{-1}$ simplifies to $\mathcal{I}^{-1}(\beta_0)$. This is the Cramer-Rao bound.

5 MLE Equals OLS in the Linear Model

▷ Handwritten Notes (what the professor said)

Maximising $\frac{1}{n}\log p(\beta, z_n)$ over $\beta \in \Theta$ is equivalent to:

$$\hat{\beta}_n = \arg\min_{\beta\in\Theta} \frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta)$$

since $-\frac{1}{2}\ln(2\pi)$ is a constant with respect to β. This is precisely the OLS estimator:

$$\hat{\beta}_n^{MLE} = \hat{\beta}_n^{OLS} = (X_n'X_n)^{-1}X_n'Y_n$$

Key Result

MLE equals OLS for the Gaussian Linear Model
Under $\varepsilon_n \mid X_n \sim N(0_{n\times 1}, I_{n\times n})$:

$$\hat{\beta}_n^{MLE} = \arg\max_{\beta\in\Theta} \frac{1}{n}\log p(\beta, z_n) = \arg\min_{\beta\in\Theta} \frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta) = \hat{\beta}_n^{OLS}$$

◇ Student's Notes

This result is specific to the Gaussian distribution. The log-likelihood is:

$$\frac{1}{n}\log p(\beta, z_n) = \underbrace{-\tfrac{1}{2}\ln(2\pi)}_{\text{constant in } \beta} - \underbrace{\tfrac{1}{2n}(Y_n - X_n\beta)'(Y_n - X_n\beta)}_{\text{OLS objective, up to sign and scale}}$$

Maximising drops the constant and flips the sign, turning the maximisation into OLS minimisation. The Gaussian is the only distribution for which the log-likelihood is a

quadratic function of the residuals. Any other distribution gives a different loss.

The Laplace connection and Lectures 7–9.
If instead $\varepsilon_{(i)} \mid X_n \overset{\text{i.i.d.}}{\sim} \text{Laplace}(0, \sigma)$, then:

$$\log p(\beta, z_n) = -n \log(2\sigma) - \frac{1}{\sigma} \sum_{i=1}^{n} |Y_{(i)} - X_{(i)}\beta|$$

Maximising over β is equivalent to minimising $\sum_i |Y_{(i)} - X_{(i)}\beta|$, which is exactly the LAD estimator from Lectures 7–9. So:

- Gaussian errors $\implies$ MLE = OLS (quadratic loss).
- Laplace errors $\implies$ MLE = LAD (absolute loss).
- General errors with density f $\implies$ MLE uses loss $-\log f(\text{residual})$.

This is why the M-estimator framework from Lectures 2–13 is so powerful: it unifies OLS, LAD, and MLE under a single template, with the choice of loss function determining which distributional assumption is being implicitly made.

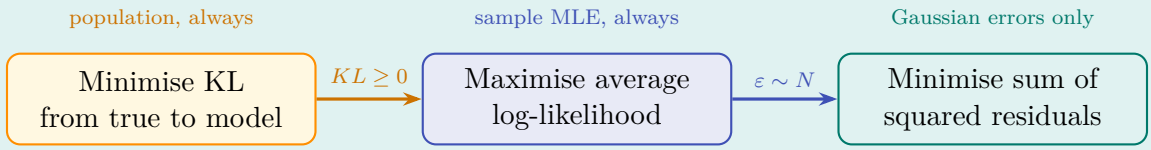

Figure 4: The logical chain. Minimising KL is always equivalent to maximising the population log-likelihood (left arrow, holds for any parametric family). Maximising the log-likelihood is equivalent to minimising squared residuals only when errors are Gaussian (right arrow). For Laplace errors the right arrow would point to absolute residuals instead.

Error distribution	Log-likelihood loss term	MLE equals
Gaussian $N(0, \sigma^2)$	$(Y_{(i)} - X_{(i)}\beta)^2/(2\sigma^2)$	OLS
Laplace$(0, \sigma)$	$\lvert Y_{(i)} - X_{(i)}\beta\rvert/\sigma$	LAD
General density f	$-\log f(Y_{(i)} - X_{(i)}\beta)$	M-estimator with loss $-\log f$

▷ Edge Cases and Pathologies

Edge Case 4: When MLE fails to be consistent.
Even in a correctly specified parametric model, MLE can fail to be consistent. The two

main pathologies are:

1. **Unbounded likelihood.** In some models the likelihood can be made arbitrarily large by a sequence of parameter values, so the argmax does not converge. The classic example is a Gaussian mixture model where one component collapses to a single data point: the variance of that component goes to zero and the likelihood goes to infinity. MLE breaks down completely. Regularisation (penalised likelihood) is needed.

2. **Growing number of parameters (incidental parameters).** If the number of parameters grows with n (e.g., a fixed effects panel model with n individual intercepts), the MLE of the structural parameter of interest (e.g., β_0) can be inconsistent even though the overall likelihood is well-behaved. This is Neyman and Scott's (1948) incidental parameters problem.

Edge Case 5: Local maxima and multimodality.
The average log-likelihood $\frac{1}{n}\sum_i \log p(\beta, z_{(i)})$ need not be concave in β for a general parametric family. It can have multiple local maxima, and the MLE defined as the global argmax may be computationally intractable to find. In practice, algorithms like EM or gradient ascent find local maxima that may not be the global one. For the linear model with Gaussian errors the log-likelihood is exactly concave (it is $-\frac{1}{2n}\|Y - X\beta\|^2$ plus a constant), so this problem does not arise. But for nonlinear models or mixture models it is a genuine concern.
Edge Case 6: Boundary solutions.
The general theory (Taylor expansion, FOC at interior point) requires $\beta_0 \in \text{Int}(\Theta)$, the same Assumption 0 from Lecture 13. If β_0 is on the boundary of Θ, the FOC $\nabla_\beta \ell_n(\hat{\beta}_n) = 0$ need not hold at the MLE. This happens, for example, when $\Theta = \{\beta : \beta_j \geq 0\}$ and the true β_0 has some component equal to zero. In that case the MLE may be a corner solution and the standard asymptotic normality result breaks down.