Lecture 4: Objective Functions and Identification

Econometrics — *continuation from Lectures 2–3*

Instructor: Prof. S. Arvanitis | **Digitisation & added notes:** T. Kourtalis

Semester: Spring 2026

▷ **Amber boxes = Handwritten Notes (professor's words)**

◇ **Teal boxes = Student's Notes**

Recall from Lectures 2–3

In Lectures 2–3 we established the semi-parametric linear model $Y_n = X_n\beta_0 + \epsilon_n$ and proved that the population objective

$$M_n^*(\beta) = (\beta_0 - \beta)' \frac{X_n' X_n}{n} (\beta_0 - \beta) + 1 \qquad (**)$$

is **uniquely minimised** at β_0 when the model is well-specified and $\text{rank}(X_n) = p$. This lecture asks: *what breaks when those conditions fail?* and *how do we go from the unobservable M_n^* to an observable estimator?*

1 Recap: The Two Pillars of Identification

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

Under certain conditions, semi-parametric statistical models have enough mathematical structure to indicate the existence of at least one objective function that is minimized at β_0.
In the linear model example, this has the form $M_n^*(\beta)$:

$$M_n^*(\beta) = (\beta_0 - \beta)' \frac{X_n' X_n}{n} (\beta_0 - \beta) + 1 \qquad (**)$$

$\arg\min_{\beta\in\Theta} M_n^*(\beta)$ uniquely minimizes at β_0, and the following played a special role in this:

i) **Well-specified model:** $Y_n = X_n\beta_0 + \epsilon_n$ and $\beta_0 \in \Theta$

ii) **Identification:** $\text{rank}(X_n) = p$

◇ Student's Notes

These two conditions do very different jobs:

Condition	What it guarantees
Well-specified	The truth β_0 actually *lives inside* Θ, so the minimum of the bowl is attainable.
Full rank	The bowl has *no flat directions* ($X'X/n$ is positive definite), so the minimum is unique.

The next two sections explore what happens when each condition is violated. Here is the well-specified, identified case for reference:

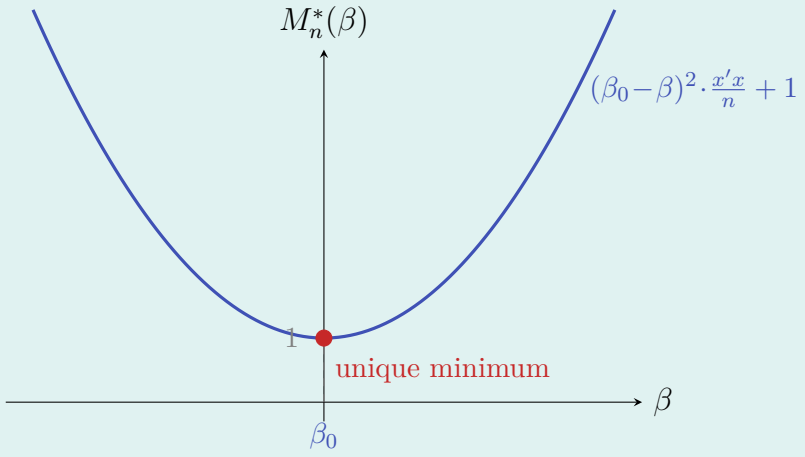

Figure 1: The "bowl" shape of $M_n^*(\beta)$ in the scalar case ($p = 1$). The unique minimum is at β_0 with value 1 (the irreducible noise variance).

2 What If Identification Fails?

▷ Handwritten Notes (what the professor said)

If $\text{rank}(X_n) < p$, then $\frac{X_n' X_n}{n}$ is positive *semi*-definite. $M_n^*(\beta)$ would have other extrema (with respect to minimization) besides β_0, and from this we would not have the information to distinguish β_0.

◇ **Student's Notes**

Why positive semi-definite ⇒ multiple minima:
If $\text{rank}(X_n) < p$, there exists a non-zero vector $v \neq 0$ in the null space of X_n (i.e. $X_n v = 0$). Then for any scalar t:

$$M_n^*(\beta_0 + tv) = (tv)'\frac{X'X}{n}(tv) + 1 = t^2 \underbrace{v'\frac{X'X}{n}v}_{=0} + 1 = 1 = M_n^*(\beta_0).$$

So the *entire line* $\{\beta_0 + tv : t \in \mathbb{R}\}$ achieves the same minimum value. The bowl has a "flat valley" along the direction v:

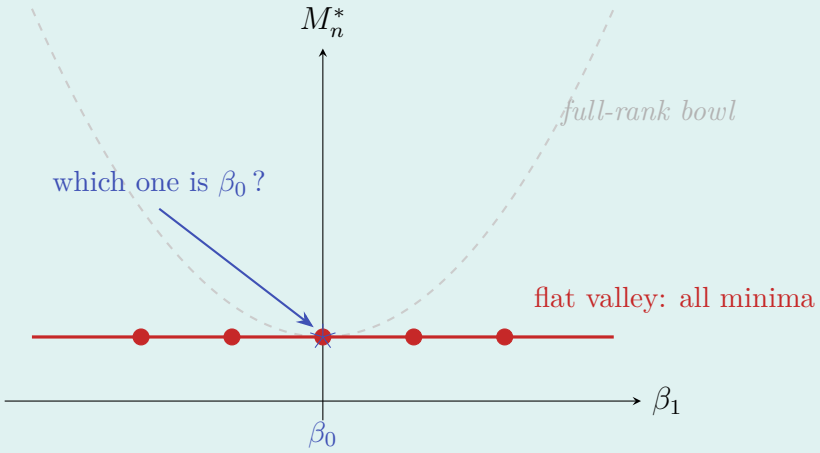

Figure 2: When $\text{rank}(X) < p$*, the bowl flattens into a valley. Every point along the valley achieves the same minimum value—*β_0 *is not identifiable. The dashed gray curve shows the full-rank case for comparison.*

Concrete example: If $x_2 = 3x_1$ (perfect multicollinearity with $p = 2$), then $\text{rank}(X) = 1 < 2$. Infinitely many (b_1, b_2) pairs give the same fitted values $X\beta$, so β_0 is not identifiable.

3 What If the Model Is Mildly Misspecified?

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

Let's examine what might happen with the properties of $M_n^*(\beta)$ if a case of "mild misspecification" holds, i.e., it holds that $Y_n = X_n\beta_0 + \epsilon_n$ but $\beta_0 \notin \Theta$.
$M_n^*(\beta)$ (as $(**)$) still has the same form and $\frac{X_n'X_n}{n}$ will be positive definite. But what happens now with $\arg\min_{\beta\in\Theta} M_n^*(\beta)$ since $\beta_0 \notin \Theta$?

[Figure: Set Θ with β_0 outside of it, and its projection to the closest point in Θ.]
In this case, and under certain conditions regarding the properties of Θ, the minimization $\arg\min_{\beta\in\Theta} M_n^*(\beta)$ will yield the unique element of Θ which is at the minimum possible distance (as shaped through M_n^*) from β_0.

◇ Student's Notes

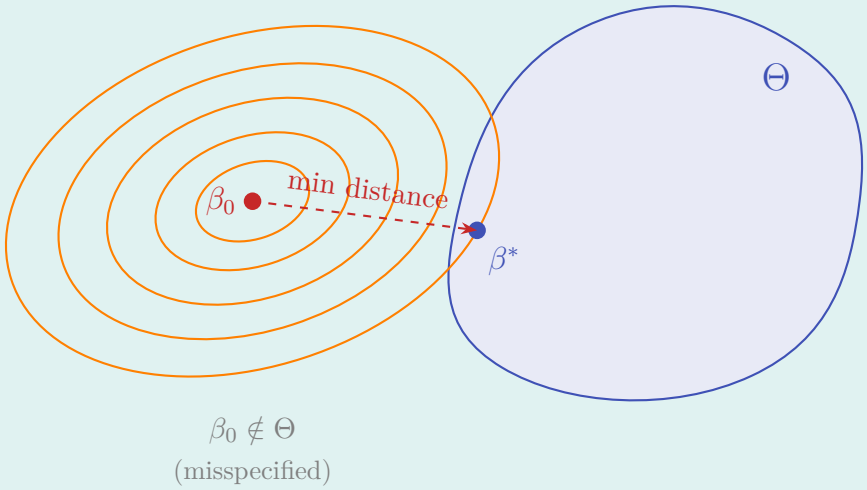

Figure 3 (Professor's figure): β_0 lies outside the parameter space Θ. The ellipses are level sets of M_n^* (contours of the bowl). The constrained minimiser $\beta^* = \arg\min_{\beta\in\Theta} M_n^*(\beta)$ is the point in Θ on the smallest contour that touches Θ—the "projection" of β_0 onto Θ under the $X'X/n$ metric.

Key takeaway: under mild misspecification the estimator still converges to *something* well-defined (β^*), but that something is *not* the truth. This is why well-specification matters so much.

Note on the distance: the "distance" here is not ordinary Euclidean distance but is warped by $X'X/n$. Directions in which the regressors have more variation are "stretched"—deviations from β_0 in those directions are penalised more heavily.

4 From Population to Sample: The Analogy Principle

▷ Handwritten Notes (what the professor said)

What we saw is that if (i) and (ii) hold, knowing M_n^* allows us to accurately find β_0 by

minimizing it:

$$\rightarrow \ (\beta_0 - \beta)' \frac{X_n' X_n}{n} (\beta_0 - \beta)$$

However, this function necessarily depends on β_0 (thus it is unobservable to us).
If we manage to find an observable function, let's say $M_n(\beta)$ (which will depend only on our sample and β), which "approximates" M_n^*, we can minimize the approximation M_n with respect to $\beta \in \Theta$, hoping that in this way we will approximate the unknown β_0.

◇ Student's Notes

This is the fundamental **logic of estimation** in this course:

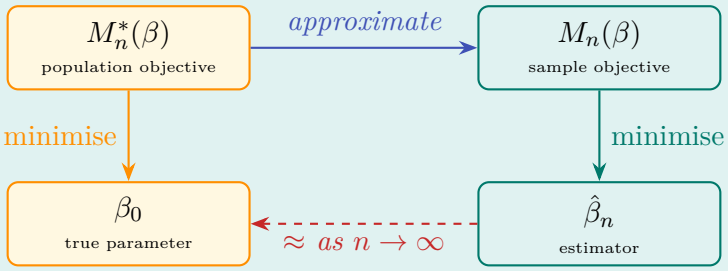

Figure 4: The estimation logic. We cannot minimise the unobservable M_n^* directly, so we minimise its sample analogue M_n instead.

The quality of the estimator $\hat{\beta}_n$ depends entirely on how well M_n approximates M_n^*. Making this "approximation" precise is the job of *consistency theory* (coming soon).

4.1 Constructing M_n for the Linear Model

▷ Handwritten Notes (what the professor said)

How could we construct M_n in this example? We recall that:

$$M_n^*(\beta) = \mathbb{E}\left[\frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta) \,\middle|\, X_n\right] = (\beta_0 - \beta)'\frac{X_n' X_n}{n}(\beta_0 - \beta) + 1$$

$\downarrow$ *Analogy Principle*
Under conditions, the function:

$$M_n(\beta) = \frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta)$$

can be an approximation of M_n^*. *(Note: This is a strong commitment; it is not the only approximation).*

◇ Student's Notes

What the Analogy Principle does: $M_n^*(\beta) = \mathbb{E}[\text{something}|X_n]$. The analogy principle says: *drop the expectation and use the "something" itself.* That is:

$$M_n^*(\beta) = \mathbb{E}\left[\frac{1}{n}\|Y - X\beta\|^2 \middle| X\right] \quad \xrightarrow{\text{drop } \mathbb{E}} \quad M_n(\beta) = \frac{1}{n}\|Y - X\beta\|^2.$$

This works because, by the law of large numbers, $M_n(\beta) \xrightarrow{p} M_n^*(\beta)$ for each β as $n \to \infty$.

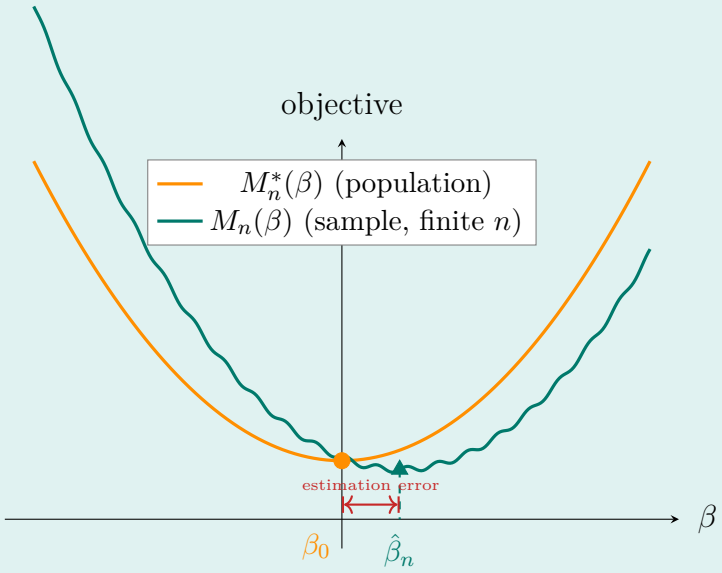

Figure 5: The sample objective M_n (teal) is a noisy, slightly shifted version of the population objective M_n^* (amber). Their minima are close but not identical for finite n. As $n \to \infty$, $M_n \to M_n^*$ and $\hat{\beta}_n \to \beta_0$.

Why "not the only approximation"? We could also use, e.g., the absolute-error loss $\frac{1}{n}\sum|y_i - x_i'\beta|$, which would give a different (LAD) estimator. Different approximations lead to different estimators with different properties.

Minimising M_n explicitly:

$$\hat{\beta}_n = \arg\min_\beta \frac{1}{n}(Y - X\beta)'(Y - X\beta)$$

Taking the first-order condition: $-\frac{2}{n}X'(Y - X\hat{\beta}) = 0$, which gives

$$\boxed{\hat{\beta}_n = (X'X)^{-1}X'Y} \qquad \text{(the OLS estimator).}$$

5 Extremum (M-) Estimators: The General Framework

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

Therefore in general:
In the (semi-)parametric models that we will study assuming they are Well-Specified, there will exist functions M_n^* that will uniquely be minimized at β_0, and which will be approximated by observable functions M_n. The minimization of these will give us a class of estimators called **Extremum Estimators (M-estimators)**.
Also, using the Estimators of M_n etc., we could construct Hypothesis Testing procedures for β_0. (We will see this later.)

Key Result

An **extremum estimator** (M-estimator) is any estimator defined as

$$\hat{\beta}_n := \arg\min_{\beta \in \Theta} \; M_n(\beta),$$

where M_n is a sample-computable objective function that approximates a population objective M_n^* whose unique minimum is at the true parameter β_0.

Estimator	$M_n(\beta)$	$\hat{\beta}_n$
OLS	$\frac{1}{n}\|Y - X\beta\|^2$	$(X'X)^{-1}X'Y$
MLE	$-\frac{1}{n}\sum \log f(z_i;\beta)$	solve score $= 0$
GMM	$\bar{g}_n(\beta)'\hat{W}\,\bar{g}_n(\beta)$	solve FOC

◇ **Student's Notes**

All the estimators in this course (OLS, MLE, GMM) are special cases of this single template. The course structure mirrors this:

1. **Identification:** show M_n^* has a unique min at β_0 (Lectures 2–4, done).
2. **Estimation:** construct M_n, define $\hat{\beta}_n = \arg\min M_n$ (this lecture).
3. **Consistency:** prove $\hat{\beta}_n \xrightarrow{p} \beta_0$ (next).
4. **Asymptotic normality:** derive the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ (later).
5. **Testing:** use the asymptotic distribution to build tests about β_0 (later still).

6 What Is an Estimator?

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

But first let's try to recall what an Estimator is (in such types of models).
For us then, an **Estimator** of β_0 will be called any function $Z_n \to \Theta$ such that it has a well-defined Probability Distribution.

Definition: Estimator

An **estimator** $\hat{\beta}_n$ of β_0 is a *measurable* function

$$\hat{\beta}_n : Z_n = (Y_n, X_n) \longrightarrow \Theta \subseteq \mathbb{R}^p,$$

i.e. a rule that maps the observed data to a point in the parameter space, and which possesses a well-defined probability distribution (induced by the randomness of the data).

◇ **Student's Notes**

Key subtleties:

- The estimator is a *random variable* (because the data are random). A specific realisation $\hat{\beta}_n = 2.7$ is called an *estimate* (non-random number).
- "Measurable" is the technical requirement that ensures probabilities like $P(\|\hat{\beta}_n - \beta_0\| > \varepsilon)$ are well-defined. For every function we encounter in this course, measurability holds automatically.
- The estimator must map into Θ, not just into $\mathbb{R}^p$. If Θ has constraints (e.g. $\sigma^2 > 0$), the estimator must respect them.

7 Asymptotic Properties: Weak Consistency

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

We will necessarily deal with the Asymptotic Properties of the Estimators (which concern their behavior as $n \to +\infty$, therefore acquiring more and more information about where β_0 is located).
A first property that will interest us is that of (weak) consistency. An estimator of β_0 will be called **weakly consistent** iff the probability of the event that the distance between the value the estimator takes and β_0 is strictly positive tends to 0.

Definition: Weak Consistency

$\hat{\beta}_n$ is **weakly consistent** for β_0 if and only if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big(\|\hat{\beta}_n - \beta_0\| > \varepsilon\big) = 0.$$

Equivalently, $\hat{\beta}_n \xrightarrow{p} \beta_0$ ("convergence in probability").

◇ Student's Notes

Unpacking the definition:

- Fix any tolerance $\varepsilon > 0$, no matter how tiny (e.g. 10^{-6}).
- Ask: "What is the probability that my estimator is more than ε away from the truth?"
- Consistency says this probability $\to 0$ as $n \to \infty$.
- Crucially, it does *not* say the estimator equals β_0 for any finite n—only that it gets *arbitrarily close* with *arbitrarily high probability* as n grows.

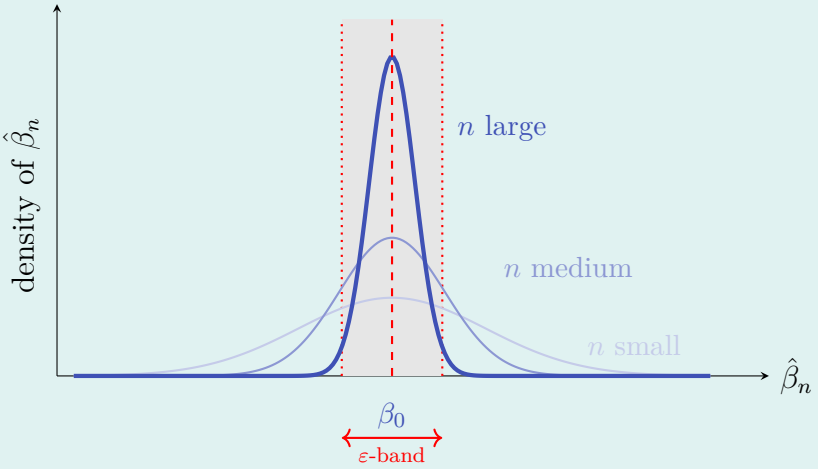

Figure 6: Convergence in probability. As n grows, the distribution of $\hat{\beta}_n$ concentrates around β_0. For any ε-band (shaded), the probability mass outside it $\to 0$.

Weak vs. strong consistency:

Type	Statement
Weak (convergence in prob.)	$P(\|\hat{\beta}_n - \beta_0\| > \varepsilon) \to 0$ for all $\varepsilon > 0$
Strong (almost sure)	$P(\hat{\beta}_n \to \beta_0) = 1$

Strong $\Rightarrow$ weak, but not conversely. In this course we mostly work with weak consistency.

Preview: to prove OLS is consistent we will need to show that $M_n(\beta) \xrightarrow{p} M_n^*(\beta)$ uniformly over Θ. The main tools will be the (weak) law of large numbers applied to $X'X/n$ and $X'\epsilon/n$.

Quick-Reference Summary

◇ Student's Notes

Lectures 2–4 narrative arc:

Lecture	What was accomplished
Lect. 2	Defined parametric vs. semi-parametric models; chose the squared-error loss
Lect. 3	Proved M_n^* has a unique minimum at β_0 (identification)
Lect. 4	Explored failures (rank deficiency, misspecification); moved from M_n^* to M_n; defined extremum estimators and weak consistency
Next	Prove $\hat{\beta}_n \xrightarrow{p} \beta_0$ (consistency of OLS)

Key new concepts this lecture:

Term	One-line meaning
Pseudo-true value β^*	Closest point in Θ to β_0 under the $X'X/n$ metric
Analogy principle	Drop the expectation from M_n^* to get M_n
Extremum estimator	$\hat{\beta}_n = \arg\min M_n(\beta)$
Weak consistency	$\hat{\beta}_n \xrightarrow{p} \beta_0$

Figures summary:

Figure	What it shows
Fig. 1	The "bowl" shape of M_n^* (well-specified + full rank)
Fig. 2	Flat valley when $\text{rank}(X) < p$ (identification failure)
Fig. 3	Projection of β_0 onto Θ under misspecification (professor's figure)
Fig. 4	The estimation logic: population → sample → estimator
Fig. 5	M_n as a noisy approximation of M_n^*
Fig. 6	Convergence in probability: distributions concentrating around β_0