Lecture 2: Foundations of Statistical Models

Econometrics — Parametric vs. Semi-Parametric Specifications

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

Semester: Spring 2026

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

◇ **Teal boxes = Student's Notes**

1 Three Dimensions of Econometric Investigation

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

When analyzing statistical models, we operate across three distinct dimensions:

1. **Methodological:** How do we find objective functions, and how do we use them to construct estimators and tests?
2. **Technical (Deriving Properties):** How do we use the properties of these objective functions to verify the behavior of our estimators (e.g., consistency, asymptotic normality)?
3. **Numerical Optimization:** How do we computationally optimize these objective functions when analytical solutions are difficult to find?

◇ **Student's Notes**

Think of these as three layers that build on each other:

Dimension	Core question	Example
Methodological	*What* to minimise?	Choose $\sum(y_i - x_i'b)^2$
Technical	*Why* does it work?	Prove $\hat{b} \xrightarrow{p} \beta_0$
Numerical	*How* to compute it?	Gradient descent, Newton–Raphson

For the linear model, we are lucky: layer 3 is trivial because OLS has a closed-form solution $(X'X)^{-1}X'Y$. For nonlinear models (logit, probit, GMM) all three layers require serious work.

2 The Basic Linear Model and Matrix Notation

▷ Handwritten Notes (what the professor said)

We observe a sample $Z_n = (Y_n, X_n)$ drawn from an unknown probability distribution D_0.
Our goal is to identify unknown characteristics of D_0.
The foundational linear model is written as:

$$Y_n = X_n \beta_0 + \varepsilon_n$$

Matrix Dimensions:

- Y_n is an $n \times 1$ random vector of the dependent variable.
- X_n is an $n \times p$ matrix of observed covariates (features).
- $\beta_0 \in \mathbb{R}^p$ is the $p \times 1$ true parameter vector we want to estimate.
- ε_n is the $n \times 1$ vector of unobserved random errors.

Core Conditional Moments:
We assume the expected value of the error, given the data, is zero:

$$\mathbb{E}(\varepsilon_n \mid X_n) = 0_{n\times 1}$$

This implies that the conditional mean of Y_n is perfectly described by our covariates:

$$\mathbb{E}(Y_n \mid X_n) = X_n \beta_0$$

Furthermore, we assume the conditional variance of the errors is a known identity matrix (assuming $\sigma^2 = 1$ for simplicity):

$$\mathbb{E}(\varepsilon_n \varepsilon_n' \mid X_n) = I_{n\times n} \quad \Longrightarrow \quad \mathrm{Var}(Y_n \mid X_n) = I_{n\times n}$$

◇ Student's Notes

Why $\mathbb{E}(\varepsilon_n|X_n) = 0$ implies $\mathbb{E}(Y_n|X_n) = X_n\beta_0$:
Apply the conditional expectation to both sides of the DGP:

$$\mathbb{E}(Y_n|X_n) = \mathbb{E}(X_n\beta_0 + \varepsilon_n|X_n) = X_n\beta_0 + \underbrace{\mathbb{E}(\varepsilon_n|X_n)}_{=0} = X_n\beta_0.$$

$X_n\beta_0$ passes through because X_n is treated as known once we condition on it.

Why $\mathbb{E}(\varepsilon_n\varepsilon_n'|X_n) = I$ implies $\mathrm{Var}(Y_n|X_n) = I$:
Since $\mathrm{Var}(\varepsilon_n|X_n) = \mathbb{E}(\varepsilon_n\varepsilon_n'|X_n) - \underbrace{\mathbb{E}(\varepsilon_n|X_n)\mathbb{E}(\varepsilon_n'|X_n)}_{=0} = I$, and $Y_n = X_n\beta_0 + \varepsilon_n$ with $X_n\beta_0$ non-random conditional on X_n, the variance of Y_n equals the variance of ε_n.

3 Defining Statistical Models

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

A statistical model is simply a *collection of possible distributions* for our sample Z_n.

Definition: Statistical Model

A statistical model $\mathcal{M}$ is a family of probability distributions $\{P_\theta : \theta \in \Theta\}$ that we consider as candidates for the true data-generating distribution D_0.

3.1 Parametric Models

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

In a parametric model, every distribution in the collection is uniquely described by a Euclidean parameter $b \in \Theta \subseteq \mathbb{R}^p$. The entire probability distribution is known, up to the parameter b.

Example: Assuming errors are exactly Normally distributed.

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

◇ **Student's Notes**

In the parametric case, once you pick b, the *entire* distribution of the data is pinned down—shape, tails, skewness, everything. This is powerful (you can write down a likelihood and use MLE) but *fragile*: if the true errors are not Normal, the model is misspecified.

Key consequence: because the distribution is fully specified, there is a **one-to-one** map

$$\beta \longleftrightarrow P_\beta \qquad \text{(parameter} \leftrightarrow \text{distribution)}.$$

3.2 Semi-Parametric Models

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

In a semi-parametric model, the model still depends on a Euclidean parameter $b \in \Theta$, but *multiple* distributions in the model can share the same parameter b. We only specify certain moments (like mean and variance) without forcing a specific shape (like the Normal distribution) on the data.

Example: The collection of all distributions for $Z_n = (Y_n, X_n)$ that simply satisfy a conditional mean of $X_n\beta$ and a variance of $I_{n\times n}$.

◇ **Student's Notes**

The map is now **many-to-one**:

$$\{P_1, P_2, P_3, \dots\} \longrightarrow \beta \qquad \text{(many distributions} \to \text{same parameter)}.$$

All distributions that share the same first two conditional moments map to the same β, regardless of their higher moments (skewness, kurtosis, etc.).

Practical implication: we *cannot* write a unique likelihood (since we don't know the full distribution), so MLE is off the table. Instead we use **moment-based** methods—OLS being the simplest.

Feature	Parametric	Semi-Parametric
What is specified	Full distribution	Selected moments only
$b \leftrightarrow P$ mapping	One-to-one	Many-to-one
Typical estimator	MLE	OLS / GMM
Robustness	Fragile to dist. shape	Robust to dist. shape
Efficiency	Most efficient *if correct*	Less efficient but safer

4 Well-Specified vs. Misspecified Models

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

- **Well-Specified:** If the true, unknown distribution D_0 is actually contained within our proposed statistical model. Here, finding the characteristics of D_0 is entirely equivalent to finding the true unknown parameter β_0.
- **Misspecified:** If D_0 is outside our model space. (For most of this course, we will rely on the strong assumption that our model is well-specified).

◇ **Student's Notes**

Visually:

Well-specified: $D_0 \in \mathcal{M}$ ✓ (truth is inside our model family)
Misspecified: $D_0 \notin \mathcal{M}$ × (truth is outside; β_0 may not even exist)

Under misspecification, the OLS estimator still converges to *something*—the "pseudo-true" value that minimises the population loss—but this pseudo-true value generally $\neq \beta_0$ and has no causal interpretation.

The well-specified assumption is what lets us say "$\arg\min M_n^*(b) = \beta_0$" in Lecture 3. Without it, the entire identification argument breaks down.

! Watch Out

Well-specification is **untestable** in a strict sense: you can test whether certain implications of the model hold in the data (e.g. Ramsey RESET test, Hausman test), but you can never prove the model is exactly correct. Every result in this course is *conditional* on this assumption.

5 The Semi-Parametric Linear Model and Optimization

▷ Handwritten Notes (what the professor said)

To find β_0 in a well-specified semi-parametric model, we need an objective function. We define the semi-parametric linear model as:

$$Y_n = X_n\beta + \varepsilon_{nn}$$

With the strict assumptions:

1. $\mathbb{E}(\varepsilon_{nn} \mid X_n) = 0_{n\times 1}$
2. $\mathbb{E}(\varepsilon_{nn}\varepsilon_{n_n}' \mid X_n) = I_{n\times n}$
3. $\text{rank}(X_n) = p \implies n \geq p$

Given this structure, we can formulate an expected objective function:

$$\mathbb{E}\left[\frac{1}{n}(Y_n - X_n\beta)'(Y_n - X_n\beta) \,\middle|\, X_n\right]$$

As proven in Lecture 3, under these exact assumptions, this function is minimized uniquely at the true parameter β_0.

◇ Student's Notes

Connecting Lectures 2 and 3:
This section sets up the *what* (the objective function); Lecture 3 proves the *why* (it has a unique minimum at β_0). The logical chain is:

1. **Lecture 2 (here):** Define the model, choose the squared-error loss.
2. **Lecture 3:** Show $M_n^*(\beta) = (\beta_0 - \beta)'\frac{X'X}{n}(\beta_0 - \beta) + 1$, so $\arg\min = \beta_0$ (identification).

3. **Next:** Replace the population objective with the sample objective $\hat{M}_n(\beta) = \frac{1}{n}(Y_n - X_n)'(Y_n - X_n\beta)$ and minimise it $\Rightarrow$ OLS estimator $\hat{\beta}_n = (X_n'X_n)^{-1}X_n'Y_n$ (estimation).

Why squared error? It is not the only possible loss function. Alternatives include absolute error (leads to LAD/median regression) or check functions (leads to quantile regression). Squared error is chosen because:

- It is differentiable everywhere, giving clean first-order conditions.
- Under the moment assumptions above, it identifies β_0 exactly.
- Under Normality, it coincides with the MLE (so nothing is lost).

Key Result

The three assumptions—exogeneity, spherical errors, and full rank—are the minimal conditions under which the population squared-error loss has a **unique global minimum** at β_0.

- Drop **exogeneity** $\Rightarrow$ the cross-term does not vanish $\Rightarrow$ bias.
- Drop **spherical errors** $\Rightarrow$ OLS still identifies β_0 but is no longer efficient (use GLS).
- Drop **full rank** $\Rightarrow$ $X'X$ is singular $\Rightarrow$ infinitely many minimisers $\Rightarrow$ no identification.

Quick-Reference: Lecture 2 at a Glance

◇ **Student's Notes**

Concept	One-line summary
Statistical model	A family of candidate distributions for the data
Parametric	Full distribution specified; $\beta \leftrightarrow P_\beta$ is one-to-one
Semi-parametric	Only some characteristics (e.g. some moments) specified; many distributions per β
Well-specified	The truth D_0 lives inside the model family
Objective function	$\mathbb{E}\left[\frac{1}{n}\|Y_n - X_n\beta\|^2 \,\middle|\, X_n\right]$; minimised at β_0
Three assumptions	Exogeneity, spherical errors, full rank