Econometrics 2

Lecture 2: Foundations of Statistical Models

Parametric vs. Semi-Parametric Specifications

AUEB | Spring Semester 2026
Taught by: Prof. S. Arvanitis

1. Three Dimensions of Econometric Investigation

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?

2. The Basic Linear Model and Matrix Notation

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:

$$Y_n = X_n b_0 + \epsilon$$

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).
- $b_0 \in \mathbb{R}^p$ is the $p \times 1$ true parameter vector we want to estimate.
- ϵ 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:

$$\mathbb{E}(\epsilon|X_n) = 0_{n\times 1}$$

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

$$\mathbb{E}(Y_n|X_n) = X_n b_0$$

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

$$\mathbb{E}(\epsilon\epsilon'|X_n) = I_{n\times n} \implies Var(Y_n|X_n) = I_{n\times n}$$

3. Defining Statistical Models

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

Parametric Models

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.
> $\epsilon|X_n \sim N(0_{n\times 1}, I_{n\times n}) \implies Y_n|X_n \sim N(X_n b, I_{n\times n})$

Semi-Parametric Models

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 $X_n b$ and a variance of $I_{n\times n}$.

4. Well-Specified vs. Misspecified Models

- **Well-Specified:** If the true, unknown distribution D_0 is actually contained within our proposed statistical model.
- **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).

5. The Semi-Parametric Linear Model and Optimization

To find b_0 in a well-specified semi-parametric model, we need an objective function. We define the semi-parametric linear model as:

$$Y_n = X_n\beta + \epsilon_n$$

With the strict assumptions:

1. $\mathbb{E}(\epsilon_n|X_n) = 0_{n\times 1}$
2. $\mathbb{E}(\epsilon_n\epsilon_n'|X_n) = I_{n\times n}$
3. $rank(X_n) = p \implies n \geq p$

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 it will be proven in Lecture 3, under these exact assumptions, this function is minimized uniquely at the true parameter b_0.