Introduction: from econometrics to machine learning

Lecture 1

Elodie Chervin Daniel Barbosa

TT 2026

Department of Economics

Course Overview

Course roadmap

1. **Wk 1** [EC]: Regression & Regularisation *(today)*
2. **Wk 2** [EC]: Classification & Validation
3. **Wk 3** [DB]: Trees & Ensembles
4. **Wk 4** [DB]: Unsupervised Learning
5. **Wk 5** [DB]: Causal ML
6. **Wk 6** [DB]: Text as Data & NLP
7. **Wk 7** [EC]: Deep learning foundations
8. **Wk 8** [EC]: LLMs and AI in research

[EC] Elodie Chervin [DB] Daniel Barbosa

Assessment

- **Weekly Questions (40%):** Conceptual and practical questions each week to consolidate learning.
- **Presentations (40%):** 10-minute presentations in Weeks 5 to 8 (2 students per week), applying methods to a real dataset. Must be accompanied by a reproducible Python script.
- **Report (20%):** A written report submitted alongside the presentation, including all Python code.

Course objectives

You will be able to implement, evaluate, and critically interpret models in an economic research context and write clean, reproducible Python code.

Why Are We Here?

The Predictive Turn in Economics

1. Structural Analysis

Workers differ in education, experience, seniority.

What is the return to a year of schooling?

→ **Inference**: causal effect.

2. Policy Targeting

A bank observes 10,000 customers and who defaulted.

Who is most likely to default next month?

→ **Prediction**: categorical outcome.

3. Automated Decisions

A firm screens CVs automatically, trained on historical hires.

Who should we interview?

→ **Algorithmic Screening**.

These are different questions, but they share a common mathematical core: learning a function that maps inputs X to an output Y.

Statistical learning: A set of tools for modelling the relationship between inputs X and an output Y by estimating the unknown function f in $Y = f(X) + \varepsilon$.

The Central Framework: $Y = f(X) + \varepsilon$

Suppose we observe a response Y and p input variables $X = (X_1, \ldots, X_p)$. We assume:

$$Y = f(X) + \varepsilon$$

- $f(X)$ is the **signal**: the systematic information X provides about Y.
- ε is the **irreducible error**: random noise, independent of X, with $\mathbb{E}[\varepsilon] = 0$.
- No matter how good our model, we can never eliminate ε.

Econometrics: understand *how* X affects Y, i.e. learn the shape and significance of f.

Machine Learning: produce $\hat{Y} = \hat{f}(X)$ that predicts as accurately as possible on new data.

Statistical learning

Supervised and unsupervised learning

Supervised learning

We observe both inputs X and a labelled output Y.

Goal: learn $\hat{f}$ so that $\hat{f}(X) \approx Y$.

Examples:

- Predict salary from education & experience *(regression)*
- Predict credit default: yes or no *(classification)*

Unsupervised learning

We observe inputs X only. No output Y exists.

Goal: discover hidden structure.

Examples:

- Group countries by economic similarity *(clustering)*
- Compress 50 survey items into 2 dimensions *(PCA)*

Supervised learning: A setting where the algorithm is trained on labelled pairs (x_i, y_i) to map new inputs to outputs.

Parametric and non-parametric

Define “parametric” and “non-parametric” models within a statistical context. Why might an economist actively *choose* a rigid parametric model over a more flexible non-parametric one?

Parametric and non-parametric models

Parametric

Assume a **fixed functional form** for f and estimate its parameters.

Example: Linear regression assumes

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Pros: interpretable, testable, requires less data.
Con: rigid; potentially biased if the true f is non-linear.

Non-Parametric

Make **no prior assumption** on the shape of f.

Examples: k-Nearest Neighbours, Random Forests, Splines.

Pros: flexible, captures complex patterns.
Cons: harder to interpret, needs more data.

Economists often choose parametric models: they impose discipline, coefficients have economic meaning, and they are easier to falsify.

Inference and prediction

Two goals

Econometrics

- Goal: *causal identification* and *inference*
- Starts from economic theory
- Cares about unbiasedness of $\hat{\beta}$
- Reports $SE(\hat{\beta})$, t-stats, p-values
- Fears endogeneity above all

Machine Learning

- Goal: *prediction* on unseen data
- Starts from data
- Cares about how well the model predicts *new, unseen* data
- Reports prediction error on a held-out test set
- Fears overfitting above all

Causal identification: Isolating the effect of X on Y from correlation.
Endogeneity: When X correlates with the error ε, biasing OLS estimates.
Inference: Drawing conclusions about a population parameter from a sample.
Overfitting: Learning the noise in training data, such that the model fails on new observations.

Prediction and inference

Explain the difference between **prediction** (forecasting $\hat{Y}$) and **inference** (understanding the causal effect of X on Y).

Provide one economic example where prediction is sufficient, and one where causal inference is required.

The Inference–Prediction Duality

Inference goal

We want $\hat{\beta}_1$ to be close to the *true parameter*.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

We study uncertainty via confidence intervals.

We care about the true effect β_1.

Prediction goal

We want $\hat{f}(x)$ to minimise errors on **data we haven't seen yet**.

$$\widehat{\mathrm{Err}}_{\text{test}} = \frac{1}{n_{\text{test}}} \sum_i (y_i - \hat{f}(x_i))^2$$

Predictions $\hat{y}$ matter more than individual parameters.

Ethics and interpretability

Consider a scenario where a firm uses an algorithm trained on historical data to automatically screen loan applications or job candidates.

What are the economic or ethical risks of deploying this model without understanding its internal logic?

Summary

- Statistical learning bridges the gap between traditional econometrics and large-scale predictive modeling.
- Our core framework is $Y = f(X) + \varepsilon$. We estimate f to either understand mechanisms (Inference) or forecast outcomes (Prediction).
- Parametric models are the default for economists, but non-parametric models offer flexibility when data is abundant.
- **Next**: Building our baseline with Linear Regression.

Statistical Foundations: Regression

Random Variables, Expectation, Variance

Let Y be a random variable with distribution P.

- **Expectation**: $\mu = \mathbb{E}[Y] = \int y \, dP(y)$ (the "centre of mass" of Y)
- **Variance**: $\sigma^2 = \mathrm{Var}(Y) = \mathbb{E}[(Y-\mu)^2]$ (spread around the mean)
- **Conditional expectation**: the best predictor of Y given X:

$$\mathbb{E}[Y \mid X = x] = \arg\min_{g} \mathbb{E}\left[(Y - g(X))^2\right]$$

  The CEF minimises mean squared error among *all* functions of X.
- **Covariance**: $\mathrm{Cov}(X, Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$
- **Law of iterated expectations**: $\mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}[Y \mid X]]$

Conditional Expectation Function (CEF): $\mathbb{E}[Y \mid X = x]$: the expected value of Y for a given value of X. It is the theoretical best predictor when minimising squared errors.

Suppose the *data generating process* (DGP) for observation i is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where x_{ij} are the **features** (or regressors) and ε_i is random noise.

Gauss-Markov assumptions:

1. **Linearity**: The true f is linear in its parameters.
2. **No perfect multicollinearity**: No feature is a perfect linear combination of the others.
3. **Homoskedasticity**: The noise ε_i has the same variance for all i.
4. **No endogeneity**: Features are uncorrelated with the noise, $\mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0$.

Under these conditions, OLS is the **Best Linear Unbiased Estimator** (BLUE): Gauss-Markov Theorem.

OLS Estimator: Bivariate Derivation

Consider the simplest case: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Ordinary Least Squares chooses $(\hat{\beta}_0, \hat{\beta}_1)$ to minimise:

$$\text{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Setting partial derivatives to zero:

- $\dfrac{\partial \text{RSS}}{\partial \beta_0} = 0 \implies \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\dfrac{\partial \text{RSS}}{\partial \beta_1} = 0 \implies \hat{\beta}_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \dfrac{\text{Cov}(x, y)}{\text{Var}(x)}$

Ordinary Least Squares (OLS): A method for estimating linear regression parameters by minimising the sum of squared residuals $\sum_i (y_i - \hat{y}_i)^2$.

OLS Geometry

Using ISL Chapter 3.4, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

OLS Geometry: The Line Always Passes Through $(\bar{x}, \bar{y})$

Claim: The fitted OLS line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ always passes through the point of means $(\bar{x}, \bar{y})$.

Proof (one line): substitute $x = \bar{x}$ into the fitted line and use the formula for $\hat{\beta}_0$:

$$\hat{y}(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \underbrace{\left(\bar{y} - \hat{\beta}_1 \bar{x}\right)}_{\hat{\beta}_0} + \hat{\beta}_1 \bar{x} = \bar{y}. \qquad \checkmark$$

Intuition: The OLS normal equation $\frac{\partial \text{RSS}}{\partial \beta_0} = 0$ is equivalent to requiring $\sum_i \hat{\varepsilon}_i = 0$, i.e. residuals sum to zero. This forces the line to pivot around the sample mean $(\bar{x}, \bar{y})$.

This is *not* a coincidence, it is a direct algebraic consequence of minimising squared errors and is an important sanity check for any OLS fit.

Evaluating Model Fit

Loss Functions: Quantifying Error

A **loss function** $\mathscr{L}(y, \hat{y})$ measures the cost of a prediction error.

- **Squared error (L_2):** $\mathscr{L}(y, \hat{y}) = (y - \hat{y})^2$. Leads to Mean Squared Error (MSE).
- **Absolute error (L_1):** $\mathscr{L}(y, \hat{y}) = |y - \hat{y}|$. Leads to Mean Absolute Error (MAE).
- **Cross-entropy** (for classification):

$$\mathscr{L}(y, \hat{p}) = -\big[y \log \hat{p} + (1 - y) \log(1 - \hat{p})\big]$$

The choice of loss function encodes our beliefs about the cost of errors. Squared errors penalise large mistakes heavily, while absolute errors treat mistakes proportionally.

Loss function: A function $\mathscr{L}(y, \hat{y})$ that quantifies the distance between a prediction and the true value.

The Challenge of High Dimensions

Multiple Regression: The Interpretation Limit

In a multivariate model, we interpret $\hat{\beta}_j$ as the marginal effect of X_j "holding all other variables constant":

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

The problem of dimensionality:

- When p is small, predictors are often distinct and interpretable.
- When p is large, predictors are inevitably correlated (multicollinearity). "Holding all else constant" becomes a logical abstraction rather than a physical reality: we cannot shift one feature without others moving in the background.

Predictive Failure: Overfitting in High Dimensions

Traditional econometrics assumes $n \gg p$. In modern contexts, we often face $p \approx n$.

- **Data Scarcity**: As p approaches n, we lose the "buffer" of observations. OLS uses its mathematical flexibility to interpolate the noise in the training sample.
- **Saturated Models**: When the number of parameters p equals the number of observations n, OLS will achieve a perfect $R^2 = 1$ even on pure random noise.
- **Result**: The model effectively "memorises" the training sample rather than learning the general population process.

Estimator Instability: Why Variance Explodes

Why does having too many variables ($p \to n$) make OLS unreliable? There are two main drivers of this instability:

- **Noise Overfitting**: With many parameters, OLS has enough "freedom" to fit the unique errors and outliers in your specific sample. It mistakes these random fluctuations for real patterns.
- **Multicollinearity**: As we add more variables, they inevitably correlate and overlap. OLS cannot distinguish their separate effects. It may give one variable a massive positive coefficient and another a massive negative one to compensate.

The Root of the Problem

Together, these make your results extremely sensitive: a tiny change in one data point can flip the coefficients entirely → High variance.

The Bias-Variance Decomposition: Derivation I

We want to predict y at a new point x^*. The true process is $y^* = f(x^*) + \varepsilon$. We minimize the **Expected Prediction Error** (EPE) at x^*:

1. **Setup**: We take the expectation $\mathbb{E}[\cdot]$ over all possible training sets.
2. **Expansion**:
$$\mathbb{E}\big[(y^* - \hat{f})^2\big] = \mathbb{E}\big[(f + \varepsilon - \hat{f})^2\big]$$
3. **Grouping**: Group the "reducible" part $(f - \hat{f})$ and the noise (ε):
$$= \mathbb{E}\big[(f - \hat{f})^2 + \varepsilon^2 + 2(f - \hat{f})\varepsilon\big]$$
4. **Independence**: Since $\mathbb{E}[\varepsilon] = 0$, the cross-term vanishes.
$$= \underbrace{\mathbb{E}\big[(f - \hat{f})^2\big]}_{\text{Reducible Error}} + \underbrace{\sigma^2}_{\text{Noise}}$$

The Bias-Variance Decomposition: Derivation II

We decompose the **Reducible Error** by adding and subtracting $\mathbb{E}[\hat{f}]$:

5. **The Trick**: Let $a = f - \mathbb{E}\hat{f}$ and $b = \mathbb{E}\hat{f} - \hat{f}$.

$$\mathbb{E}\left[(f - \hat{f})^2\right] = \mathbb{E}\left[(a+b)^2\right] = \mathbb{E}[a^2 + b^2 + 2ab]$$

6. **Expanding the Square**:
   - $\mathbb{E}[a^2] = (f - \mathbb{E}\hat{f})^2 \quad \rightarrow$ **Bias**^{2} (since a is constant)
   - $\mathbb{E}[b^2] = \mathbb{E}\left[(\mathbb{E}\hat{f} - \hat{f})^2\right] \quad \rightarrow$ **Variance**
   - $2\mathbb{E}[ab] = 2a\mathbb{E}[b] = 2a(\mathbb{E}\hat{f} - \mathbb{E}\hat{f}) = 0$

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \sigma^2$$

Interpreting the Components

- **Bias**^{2}: The error from approximating a complex process with a model that is "too simple." A high-bias model ignores the signal (underfitting).
- **Variance**: The error from a model being "too sensitive" to the specific training data. A high-variance model mistakes noise for signal (overfitting).
- **Irreducible Noise** (σ^2): The fundamental uncertainty in the world.

Minimising total error requires a balance. To reduce variance, we must often accept some bias by restricting the model's complexity.

Bias-variance tradeoff: The inverse relationship between model bias and model variance; reducing one typically increases the other.

From Unbiasedness to Regularisation

Econometric theory traditionally prioritises **Unbiasedness** (Bias = 0). However, the decomposition reveals that this is not always optimal for prediction.

- In high dimensions, an unbiased OLS estimator has such extreme variance that its expected error is massive.
- We can achieve a lower **Total Error** by deliberately introducing a small amount of bias.
- By "shrinking" our estimates, we sacrifice unbiasedness to gain a substantial reduction in variance.

A slightly biased model that is stable is often more accurate than an unbiased model that is unstable.

Summary

- In high dimensions, OLS becomes unstable and captures noise (overfitting).
- The Bias-Variance Tradeoff is the core constraint of statistical learning.
- Minimising Total Error is the goal, which may require accepting some bias.
- **Regularisation**: A systematic method for trading bias to reduce estimator variance.

Regularisation and Shrinkage

The Regularisation Idea

Introduce a **penalty** for coefficient magnitude to the objective function:

$$\text{Minimise:} \quad \text{RSS} + \lambda \cdot \text{Penalty}(\beta)$$

$\lambda \geq 0$ is the **tuning parameter** (higher λ results in a simpler model).

Ridge Regression (L_2 Regularisation)

Ridge penalises the **sum of squared coefficients**:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- **Shrinkage**: Pulls all β_j toward zero, but never *exactly* zero.
- **Stability**: Prevents coefficients from exploding in the presence of correlated variables.
- **Scaling**: **Crucial!** Because we penalise magnitude, you must standardise variables (mean 0, var 1) so the penalty is fair.

Shrinkage: A technique that trades a small amount of bias for reduction in variance by pulling coefficient estimates toward zero.

Lasso Regression (L_1 Regularisation)

Lasso penalises the **sum of absolute values**:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **Sparsity**: Unlike Ridge, Lasso forcefully slams weak coefficients to **exactly zero**.
- **Feature Selection**: It identifies which variables are useful and discards the rest.
- **Interpretation**: Results in a simpler, more interpretable model with a subset of active predictors.

Lasso (L1): Least Absolute Shrinkage and Selection Operator. A regularisation method that performs both variable shrinkage and automatic feature selection.

Ridge vs. Lasso: Comparing Shrinkage Forces

Why does the choice of penalty (L_1 vs L_2) change the outcome?

Ridge (L_2): Smooth

Penalty Slope: $2\lambda\beta$

- As $\beta \to 0$, the penalty force diminishes.
- There is no final pressure to reach exactly zero.
- **Result**: Shrinks all variables but keeps them in the model.

Lasso (L_1): Sharp

Penalty Slope: $\pm\lambda$

- The penalty force remains constant even near zero.
- Even tiny coefficients feel a full-strength push toward zero.
- **Result**: Drops weak variables by setting $\beta = 0$.

Elastic Net & Tuning λ

- **Elastic Net**: A hybrid that combines both L_1 and L_2 penalties.

$$\text{Penalty} = \lambda \left[\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 \right]$$

  Best of both worlds: selection (Lasso) + stability (Ridge).

- **How to choose λ?**: We use **Cross-Validation**.

K equal-sized "folds" (e.g., $K = 5$)

Hyperparameter: A parameter (like λ or α) whose value is set before the learning process begins, rather than being estimated from the data like β.

Synthesis: The Modern Economist's Toolkit

Econometrics

Causal Identification

Providing stable mechanisms and interpretable estimates of the "why" behind the data.

Machine Learning

Predictive Precision

Leveraging high-dimensional data to produce accurate forecasts or predictions.

Applied AI

Scale & Discovery

Processing unstructured data (text, images) to broaden the scope of economic research.

Summary

- Econometrics targets causal inference: predictive modelling targets accuracy. Both are complementary tools in an economist's kit.
- **Central framework**: $Y = f(X) + \varepsilon$. Statistical learning estimates f from data.
- **Parametric and non-parametric**: parametric models impose structure; non-parametric models are flexible.
- **Bias-variance tradeoff**: OLS is unbiased but can have high variance. Regularisation accepts bias to reduce variance.
- **Ridge** (L2): shrinks coefficients toward zero; stabilises correlated inputs.
- **Lasso** (L1): forces weak coefficients to exactly zero; performs variable selection.
- **Elastic Net**: combines L1 and L2; robust for correlated groups.
- **Next week**: classification, evaluation metrics, and cross-validation.