SOC 690S: Machine Learning in Causal Inference

Week 1: Motivation and Linear Regression

Wenhao Jiang
Department of Sociology, Fall 2025

Introduction

What to Expect in this Course

- This is an *advanced statistics course* combining *causal inference* (statistical inference) with *prediction* (machine learning)
    - Emphasis on both *statistical theory* and *sociological applications*

What to Expect in this Course

- This is an *advanced statistics course* combining *causal inference* (statistical inference) with *prediction* (machine learning)
  - Emphasis on both *statistical theory* and *sociological applications*
- This integration is a *growing frontier*
  - Driven by high-dimensional data ($p > n$)
  - Enabled by flexible, non-linear models

What to Expect in this Course

- This integration is a *growing frontier* driven by high-dimensional data
  - Driven by high-dimensional data ($p > n$)
  - Wang et al. (2024) estimated the causal (hopeful) effect of biomarkers on Alzheimer's Disease severity using high-dimensional genetic data
  - Gupta and Lee (2023) decomposed causal effects of components in digital marketing interventions, where firms track thousands of features—such as user behavior, timestamps, and campaign attributes

What to Expect in this Course

- This integration is a *growing frontier* enabled by flexible, non-linear models
    - Óskarsdóttir et al. (2020) incorporated mobile phone call-detail records and social network measures—vast, nontraditional datasets—into credit scoring models, using ML methods such as random forests and gradient boosting to flexibly estimate non-linear interactions among predictors

Tips of Study

- I assume you have reasonable familiarity with *Probability and Statistics*, and a basic understanding of *Calculus* and *Linear Algebra*
- However, you do not need to follow every step of the statistical derivations
- The focus is on developing *intuition* (for example, how and why *Double Machine Learning* works for statistical inference in high-dimensional data) and understanding how these methods may be applied in your research
    - Homework and the midterm exam are intended as learning tools to strengthen your *basic* statistics and build *intuition*

Tips of Study

- I will go through the material at a deliberate pace
- The pace and content will remain flexible, tailored to your level and needs
- Don't feel pressured if some statistical concepts are unclear at first
  - Some topics are not immediately essential
  - Others will become more familiar through repeated exposure
- My slides are intentionally dense (to help me prepare), so please feel free to stop me at any point if something is unclear

Linear Regression and Conditional Expectation Function (CEF)

Conditional Expectation Function (CEF)

- In a *population*, given a dependent variable Y_i and a $p \times 1$ vector of covariates X_i, the *best predictor* of Y_i given X_i is

$$g(X_i) = E[Y_i \mid X_i]$$

  in the sense of minimizing mean squared error (MMSE)
- X_i is a random variable, and $E[Y_i \mid X_i]$, as a function of X_i, is also a random variable
- We sometimes work with a particular value of CEF

$$E[Y_i \mid X_i = x] = \int t f_y(t \mid X_i = x) dt$$

Conditional Expectation Function (CEF)

- Suppose X_i is years of completed education, Y_i is weekly earnings

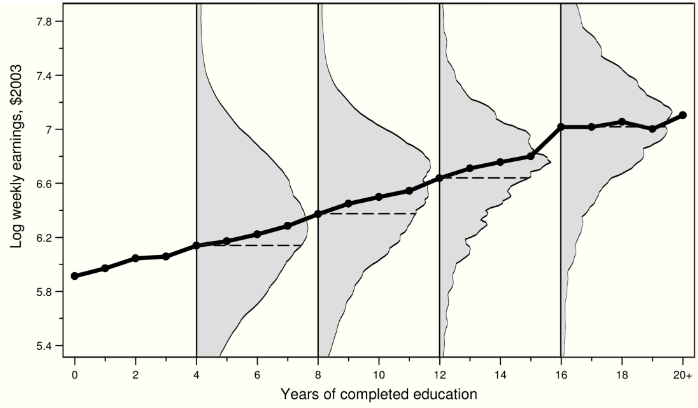

Figure: The CEF of average weekly earnings given schooling

Conditional Expectation Function (CEF)

- *Law of Iterated Expectation*

$$E[Y_i] = E[E[Y_i \mid X_i]]$$

Conditional Expectation Function (CEF)

- *CEF Decomposition Property*

$$Y_i = E[Y_i \mid X_i] + \epsilon_i$$

- ϵ_i is mean independent of X_i, that is $E[\epsilon_i | X_i] = 0$

$$E[Y_i - E[Y_i \mid X_i] \mid X_i] = E[Y_i \mid X_i] - E[Y_i \mid X_i]$$

- ϵ_i is mean independent of any function of X_i, that is $E[\epsilon_i | m(X_i)] = E[\epsilon_i m(X_i)] = 0$

$$E[\epsilon_i m(X_i)] = E[E[\epsilon_i m(X_i) \mid X_i]] = E[m(X_i)E[\epsilon_i \mid X_i]] = 0$$

- Any random variable Y_i can be decomposed into a piece that is *explained by* X_i (CEF) and a piece left over that is orthogonal to any function of X_i

Conditional Expectation Function (CEF)

- *CEF Prediction Property*
- Let $m(X_i)$ by any function of X_i, the CEF is the MMSE predictor of Y_i given X_i

$$E[Y_i \mid X_i] = \underset{m(X_i)}{\arg\min}\, E[(Y_i - m(X_i))^2]$$

$$\begin{aligned}(Y_i - m(X_i))^2 &= ((Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))^2) \\ &= (Y_i - E[Y_i \mid X_i])^2 + 2(E[Y_i \mid X_i] - \\ &\quad m(X_i))(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - m(X_i))^2\end{aligned}$$

The last term is minimized at 0 when $m(X_i)$ is the CEF

Linear Regression

- The linear regression we typically deal with—the Ordinary Least Squares (OLS)—minimizes mean squared errors
- The solution minimizing MSE, $X_i'\beta$, is the *Best Linear Predictor* (BLP)

Linear Regression

- The linear regression we typically deal with—the Ordinary Least Squares (OLS)—minimizes mean squared errors
- The solution minimizing MSE, $X_i'\beta$, is the *Best Linear Predictor* (BLP)
- At *population* level, given a $p \times 1$ covariates X_i, the $p \times 1$ regression coefficient vector β is defined by solving

$$\beta = \arg\min_b E[(Y_i - X_i'b)^2]$$

- Using the first-order condition (FOC),

$$E[-X_i(Y_i - X_i'\beta)] = 0 \longrightarrow \textit{Normal Equation}$$
$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$

Linear Regression

- The linear regression we typically deal with—the Ordinary Least Squares (OLS)—minimizes mean squared errors
- The solution minimizing MSE, $X_i'\beta$, is the *Best Linear Predictor* (BLP)
- At *population* level, given a $p \times 1$ covariates X_i, the $p \times 1$ regression coefficient vector β is defined by solving

$$\beta = \arg\min_b E[(Y_i - X_i'b)^2]$$

- Using the first-order condition (FOC),

$$E[-X_i(Y_i - X_i'\beta)] = 0 \rightarrow \textit{Normal Equation}$$
$$\beta = E[X_iX_i']^{-1}E[X_iY_i]$$

- *By construction*, the population residual defined as $e_i \equiv Y_i - X_i'\beta$ is orthogonal to X_i ($e_i \perp X_i$); $E[X_i(Y_i - X_i'\beta)] = E[X_ie_i] = 0$

Linear Regression

- The *Regression CEF Function*
- The function $X_i'\beta$ provides the MMSE linear approximation to the CEF $E[Y_i \mid X_i]$

$$\beta_{CEF} = \arg\min_b E[(E[Y_i \mid X_i] - X_i'b)^2]$$

Note that β solves $\arg\min_b E[(Y_i - X_i'b)^2]$

$$\begin{aligned} E[(Y_i - X_i'b)^2] &= E[\{(Y_i - E[Y_i \mid X_i]) + (E[Y_i \mid X_i] - X_i'b)\}^2] \\ &= E[(Y_i - E[Y_i \mid X_i])^2] + E[(E[Y_i \mid X_i] - X_i'b)^2] + \\ &\quad E[(Y_i - E[Y_i \mid X_i])(E[Y_i \mid X_i] - X_i'b)] \end{aligned}$$

- The first term is not related to β, the last term is zero from *CEF Decomposition Property*

Linear Regression

- If CEF is *linear*, then the population linear regression is it (proof omitted)
- If CEF is *nonlinear*, the population linear regression still provides the BLP (or equivalently, *Best Linear Approximation*, BLA)

Linear Regression

- Let us be a little slower for matrix operation that will pay off later
- Suppose $p = 2$

$$X_i = \begin{bmatrix} 1 \\ D_i \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad e_i \text{ and } Y_i \text{ are } \textit{scalars}$$

$$Y_i = X_i'\beta + e_i = \begin{bmatrix} 1 & D_i \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + e_i = \beta_0 + \beta_1 D_i + e_i$$

- The *Normal Equation* is

$$E\left[\begin{bmatrix} 1 \\ D_i \end{bmatrix} (Y_i - \beta_0 - \beta_1 D_i)\right] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Linear Regression

- The *Normal Equation* gives

$$E[Y_i - \beta_0 - \beta_1 D_i] = 0$$
$$E[D_i(Y_i - \beta_0 - \beta_1 D_i)] = 0$$

- This is the bivariate regression in a non-matrix form you typically see

Linear Regression

- Solving the *Normal Equation* in matrix form $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$

$$\beta = E\left[\begin{bmatrix} 1 \\ D_i \end{bmatrix} \begin{bmatrix} 1 & D_i \end{bmatrix}\right]^{-1} E\left[\begin{bmatrix} Y_i \\ D_i Y_i \end{bmatrix}\right]$$
$$= \begin{bmatrix} 1 & E[D_i] \\ E[D_i] & E[D_i^2] \end{bmatrix}^{-1} \begin{bmatrix} E[Y_i] \\ E[D_i Y_i] \end{bmatrix}$$

- For the inverse of a matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Linear Regression

- Solving the *Normal Equation* in matrix form $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$

$$\beta = \frac{1}{E[D_i^2] - E[D_i]^2} \begin{bmatrix} E[D_i^2] & -E[D_i] \\ -E[D_i] & 1 \end{bmatrix} \begin{bmatrix} E[Y_i] \\ E[D_i Y_i] \end{bmatrix}$$

$$= \frac{1}{E[D_i^2] - E[D_i]^2} \begin{bmatrix} E[D_i^2]\,E[Y_i] - E[D_i]\,E[D_i Y_i] \\ -E[D_i]\,E[Y_i] + E[D_i Y_i] \end{bmatrix}$$

- Re-arranging the terms

$$\beta_0 = \frac{E[D_i^2]\,E[Y_i] - E[D_i]\,E[D_i Y_i]}{E[D_i^2] - E[D_i]^2}$$

$$\beta_1 = \frac{E[D_i Y_i] - E[D_i]\,E[Y_i]}{E[D_i^2] - E[D_i]^2}$$

Partialling Out

Understanding β in Multivariate Regression

The $p \times 1$ vector $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$ does not give much information about each β component in a multivariate regression

Understanding β in Multivariate Regression

The $p \times 1$ vector $\beta = E[X_i X_i']^{-1} E[X_i Y_i]$ does not give much information about each β component in a multivariate regression

Suppose we have vector of regressors X_i partitioned into two components

$$X_i = (D_i, W_i')'$$

where D represents the "target" regressor of interest, and W represents the other regressors (or controls). We write

$$Y_i = \beta_1 D_i + \beta_2' W_i + e_i$$

Understanding β in Multivariate Regression

How does the predicted value of Y change if D increases by a unit, *while holding W unchanged*?

- What is the difference in predicted wages between men and women with the same characteristics of human capital?

Understanding β in Multivariate Regression

How does the predicted value of Y change if D increases by a unit, *while holding W unchanged*?

- What is the difference in predicted wages between men and women with the same characteristics of human capital?

The *Frisch-Waugh-Lovell Theorem* states that the equation is equivalent to

$$\begin{aligned}\tilde{Y}_i &= \beta_1 \tilde{D}_i + \tilde{e}_i \\ \textit{where } \tilde{D}_i &= D_i - \gamma'_{DW} W_i \\ \gamma_{DW} &= \arg\min_{\gamma} E\left[(D_i - \gamma' W_i)^2\right]\end{aligned}$$

Understanding β in Multivariate Regression

The *Frisch-Waugh-Lovell Theorem* states that the equation is equivalent to

$$\tilde{Y}_i = \beta_1 \tilde{D}_i + \tilde{e}_i$$

The estimation of β_1 is now transformed from a *multivariate* regression to a *bivariate* regression

$$\beta_1 = \arg\min_{b_1} E[(\tilde{Y}_i - b_1 \tilde{D}_i)^2]$$

Solving FOC

$$E[\tilde{D}_i(\tilde{Y}_i - \beta_1 \tilde{D}_i)] = 0 \rightarrow \beta_1 = \frac{E[\tilde{D}_i \tilde{Y}_i]}{E[\tilde{D}_i^2]}$$

β in Multivariate Regression Using a *Sample*

Suppose we have a sample analog of OLS

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 D_i + \beta_2 W_{1i} + \ldots + \beta_k W_{ki} + e_i \\ D_i &= \gamma_0 + \gamma_1 W_{1i} + \ldots + \gamma_k W_{ki} + \check{D}_i \\ \hat{\beta}_1 &= \frac{Cov(Y_i, \check{D}_i)}{V(\check{D}_i)} = \frac{Cov(\check{Y}_i, \check{D}_i)}{V(\check{D}_i)} \end{aligned}$$

Equation holds using either $\check{Y}_i$ or Y_i

β in Multivariate Regression Using a *Sample*

To show this is the case, notice that

- $\check{D}_i$ is a linear combination of all regressors, D_i and W_i', both of which are uncorrelated with e_i
- $\check{D}_i$ already partials out W_i'; $\check{D}_i \perp\!\!\!\perp W_i$
- For the same reason, $Cov(D_i, \check{D}_i) = V(\check{D}_i)$

$$\frac{Cov(Y_i, \check{D}_i)}{V(\check{D}_i)} = \frac{Cov(\beta_0 + \beta_1 D_i + \beta_2 W_{1i} + \ldots + \beta_k W_{ki} + e_i, \check{D}_i)}{V(\check{D}_i)}$$

$$= \frac{Cov(\beta_1 D_i, \check{D}_i)}{V(\check{D}_i)} = \hat{\beta}_1$$

β in Multivariate Regression Using a *Sample*

To show this is the case, notice that

- $\check{D}_i$ is a linear combination of all regressors, D_i and W_i', both of which are uncorrelated with e_i
- $\check{D}_i$ already partials out W_i'; $\check{D}_i \perp\!\!\!\perp W_i$
- For the same reason, $Cov(D_i, \check{D}_i) = V(\check{D}_i)$

$$\frac{Cov(Y_i, \check{D}_i)}{V(\check{D}_i)} = \frac{Cov(\beta_0 + \beta_1 D_i + \beta_2 W_{1i} + ... + \beta_k W_{ki} + e_i, \check{D}_i)}{V(\check{D}_i)}$$
$$= \frac{Cov(\beta_1 D_i, \check{D}_i)}{V(\check{D}_i)} = \hat{\beta}_1$$

Equation holds using either $\check{Y}_i$ or Y_i, as the part being partialled out (W_i') from Y_i is uncorrelated with $\check{D}_i$

β in Multivariate Regression Using a *Sample*

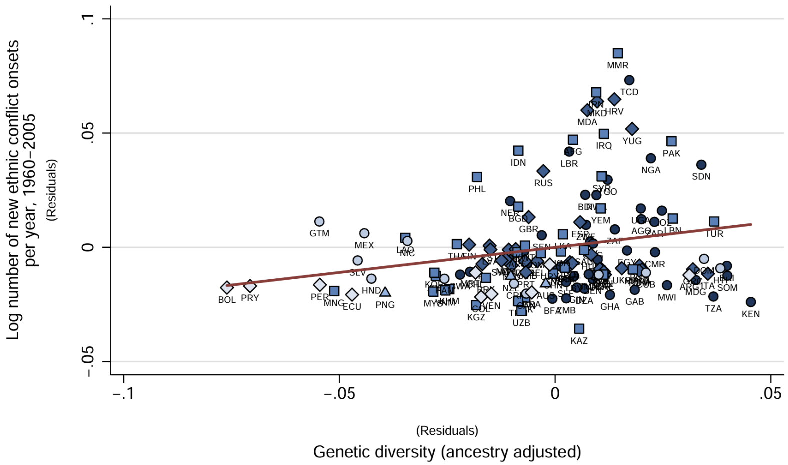

Figure: The Nature of Conflict (Arbatli, Ashraf, and Galor 2015)

Asymptotic Property of OLS

Asymptotic Property of OLS

We are interested in the distribution of the *sample* analog of

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i]$$

$$\text{where } X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix} \in \mathbb{R}^{p \times 1} \text{ and } Y_i \text{ is a } \textit{scalar}$$

Suppose $[Y_i X_i']'$ is *independently and identically distributed* in a sample of size n. The OLS estimator is given by

$$\hat{\beta} = \left[\sum_i X_i X_i' \right]^{-1} \sum_i X_i Y_i$$

Asymptotic Property of OLS

- Given $Y_i = X_i'\beta + e_i$

$$\hat{\beta} = \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i \left(X_i'\beta + e_i\right)$$
$$= \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Under regularity conditions $E||X_i||^2 < \infty$, $E\left[e_i^2||X_i||^2\right] < \infty$, $E[X_i X_i']$ is invertible ($E[X_i X_i'] \succ 0$), and $p/n \to 0$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1}\right)$$

Asymptotic Property of OLS

$\hat{\beta}$ is $\sqrt{n}$-consistent

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1}\right)$$

The consistent "sandwich" estimator (Eicker-Huber-White) of a *sample* is then given by

$$\hat{V}(\hat{\beta}) = (X_i X_i')^{-1} \left(\sum_i^n X_i X_i' \hat{e}_i^2 \right) (X_i X_i')^{-1}$$

by plugging in sample $\hat{e}_i^2$ to estimate e_i^2

This is also known as heteroskedasticity-consistent standard errors (*robust*).

Asymptotic Property of OLS

- This is, however, not the standard error you get by default from packaged software.
- Default standard errors are derived under a homoskedasticity assumption $E[e_i^2|X_i] = \sigma^2$
- Given the assumption, we have the "meat"

$$E[e_i^2 X_i X_i'] = E[E[e_i^2 X_i X_i'|X_i]] = \sigma^2 E[X_i X_i']$$

Accordingly,

$$\begin{aligned} E[X_i X_i']^{-1} E[e_i^2 X_i X_i'] E[X_i X_i']^{-1} &= \sigma^2 E[X_i X_i']^{-1} E[X_i X_i'] E[X_i X_i']^{-1} \\ &= \sigma^2 E[X_i X_i']^{-1} \end{aligned}$$

Asymptotic Property of OLS

- When p/n is not small, the "sandwich" estimate becomes inconsistent and underestimated
- The last chapter of MHE discusses the issue in detail, and here I give the intuition
- when $p/n \to c > 0$, the *operator norm error* no longer vanishes, but grows at rate of $\sqrt{p/n}$

$$\left\| \frac{1}{n} X_i X_i' - E[X_i X_i'] \right\|_{\text{op}} = \sup_{\|v\|_2=1} \left| v' \left(\frac{1}{n} X_i X_i' - E[X_i X_i'] \right) v \right| \sim O_p(\sqrt{\frac{p}{n}})$$

- The intuition is that each entry of $\frac{1}{n} X_i' X_i$ still satisfies LLN; however there are $p \times p$ entries. Ensuring all of them to be consistent is much harder, and the LLN fails in operator norm.

Neyman Orthogonality

β in Multivariate Regression Using a *Sample*

Adaptive Statistical Inference

- Under regularity conditions and if $p/n \approx 0$, the estimation error in $\check{D}_i$ and $\check{Y}_i$ has no first-order effect on the stochastic behavior of $\hat{\beta}_1$

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} \mathcal{N}\left(0, E[\tilde{D}^2]^{-1}E[\tilde{D}^2e^2]E[\tilde{D}^2]^{-1}\right)$$

- Note the sample estimate of $\hat{V}(\hat{\beta}_1)$ is the same heteroskedasticity robust standard errors we derived before

Adaptive Statistical Inference

- The *Adaptive Statistical Inference* points to the fact that estimation of residuals $\check{D}$ has a negligible impact on the large sample behavior of the OLS estimate
- The approximate behavior is the same as if we had used true residuals $\check{D}$ instead

From FWL to Neyman Orthogonality (Quick Summary)

- The *adaptivity* property will be derived later as a consequence of a more general phenomenon called *Neyman orthogonality*
- Formally,

$$\frac{\partial}{\partial \eta} E[\psi(Z;\theta,\eta)]\bigg|_{\eta=\eta_0} = 0$$

- where Z is the observed data, θ is the target parameter, η is the estimated nuisance function, and η_0 is the true nuisance function
- $\psi(\cdot)$ is the score function; in the OLS case, it is the normal equation

From FWL to Neyman Orthogonality (Quick Summary)

Neyman Orthogonality

$$\frac{\partial}{\partial \eta} E[\psi(Z;\theta,\eta)]\Big|_{\eta=\eta_0} = 0$$

When *Neyman Orthogonality* is satisfied (OLS satisfies it by design)

- Small errors in estimating η (that affects moment only at second order) does not change the fact that one still get $\sqrt{n}$-consistent, asymptotically normal inference for θ
- One can more flexibly estimate η, even in the case of non-linearity and high-dimensional data, using machine learning (ML)
- This is one of the key motivations of Double Machine Learning (DML)