The Math of Linear Regression:

Reintroducing an Old Acquaintance

Mathematics for Political Scientists

Carlos Gueiros

Fall 2025

University of Mannheim

Simple Linear Regression (Scalar Form)

For observations indexed by $i = 1, \ldots, n$, consider the model

$$Y_i \;=\; \beta_0 + \beta_1 X_i + u_i.$$

- Y_i is the dependent variable; X_i is the independent variable; u_i is the disturbance.
- The least squares estimates $(\hat{\beta}_0, \hat{\beta}_1)$ minimize the sum of squared residuals

$$S(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_i\big)^2.$$

- First-order conditions:

$$\sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_i\big) = 0, \qquad \sum_{i=1}^{n} X_i \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0.$$

- Closed-form solution:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \, \bar{X}.$$

Multiple Linear Regression (Scalar Form)

With k regressors, the model for each observation $i = 1, \ldots, n$ is

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i.$$

- The least squares estimates $\hat{\beta}_0, \ldots, \hat{\beta}_k$ minimize

$$S(\beta_0, \ldots, \beta_k) \;=\; \sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\big)^2.$$

- First-order conditions (one equation per parameter):

$$\sum_{i=1}^{n} \big(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\big) = 0,$$

$$\sum_{i=1}^{n} X_{ij} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\right) = 0, \quad j = 1, \ldots, k.$$

- These $k + 1$ linear equations motivate the *Matrix Form*, where the system is written compactly and solved using linear algebra.

The True Model (Matrix Form)

Our statistical model will essentially look something like the following:

$$\underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}_{n \times 1} = \underbrace{\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}}_{n \times (k+1)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{(k+1) \times 1} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{n \times 1}$$

The True Model (Matrix Form)

This can be rewritten more simply as:

$$y = X\beta + \varepsilon \tag{1}$$

This is assumed to be an accurate reflection of the real world.

The model has a *systematic component* ($X\beta$) and a *stochastic component* (ε). Our goal is to obtain estimates of the population parameters in the vector β.

Criteria for Estimates

Our *estimates* of the population parameters are denoted by $\hat{\beta}$. The criterion we use is to find the estimator $\hat{\beta}$ that minimizes the *sum of squared residuals*.

The vector of residuals is

$$e = y - X\hat{\beta}. \tag{2}$$

Note: Carefully distinguish between *disturbances* ε (unobserved) and *residuals* e (observed). In general, $\varepsilon \neq e$.

Sum of Squared Residuals

The sum of squared residuals (RSS) is $e'e$.

$$\underbrace{\begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix}}_{1\times n} \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}}_{n\times 1} = \underbrace{\begin{bmatrix} e_1^2 + e_2^2 + \cdots + e_n^2 \end{bmatrix}}_{1\times 1}.$$

It follows that

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) && (4) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}, && (1) \end{aligned}$$

where we used that the transpose of a scalar equals itself, so $(y'X\hat{\beta})' = \hat{\beta}'X'y$.

Differentiating the RSS

To find the $\hat{\beta}$ that minimizes $e'e$, differentiate (4) with respect to $\hat{\beta}$:

$$\frac{\partial (e'e)}{\partial \hat{\beta}} \;=\; -2X'y \;+\; 2X'X\,\hat{\beta} \;=\; 0. \tag{5}$$

To check this is a minimum, differentiate again with respect to $\hat{\beta}$:

$$\frac{\partial^2 (e'e)}{\partial \hat{\beta}\,\partial \hat{\beta}'} \;=\; 2\,X'X,$$

which is positive definite when X has full column rank.

Normal Equations

From (5) we obtain the *normal equations*:

$$(X'X)\,\hat{\beta} \;=\; X'y. \tag{10}$$

Two immediate facts about $X'X$:

- It is always square ($k \times k$).
- It is symmetric.

Closed-Form Estimator

Since $X'X$ and $X'y$ are known from the data while $\hat{\beta}$ is unknown, if $(X'X)^{-1}$ exists, premultiply both sides of (10) by the inverse:

$$(X'X)^{-1}(X'X)\,\hat{\beta} \;=\; (X'X)^{-1}X'y. \tag{11}$$

Because $(X'X)^{-1}(X'X) = I_k$, this gives

$$\hat{\beta} \;=\; (X'X)^{-1}X'y. \tag{12}$$

Up to this point, no stochastic assumptions were required. The OLS estimators are linear functions of the observed data (X, y).

Model and Assumptions

Model (definition): $y = X\beta + \varepsilon$, with β unknown.

What requires assumptions: the stochastic behavior of ε (and how rows of (X, y) are generated).

We work *conditional on* X: probabilities and expectations are taken given the realized design matrix X.

Design Matrix X

- X is treated as *given* (nonstochastic) for estimation and inference; if X is random, we condition on its realization.
- **Full column rank:** $\operatorname{rank}(X) = k$ with $k < n$. Then $X'X$ is invertible and the normal equations have a unique solution

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Zero Conditional Mean (Exogeneity)

$$\mathbb{E}[\varepsilon \mid X] = 0.$$

Interpretations:

- Given X, the disturbance has mean zero for every observation: regressors carry no information about the mean of ε.
- No omitted component correlated with any column of X.

Consequence: $\mathbb{E}[\hat{\beta} \mid X] = \beta$ (unbiasedness).

Error Covariance: Spherical (CLM)

$$\mathbb{E}[\varepsilon\varepsilon' \mid X] = \sigma^2 I_n.$$

This encodes *homoskedasticity* and *no autocorrelation*:

$$\mathbb{E}(\varepsilon_i^2 \mid X) = \sigma^2 \quad \text{for all } i, \qquad \mathbb{E}(\varepsilon_i \varepsilon_j \mid X) = 0 \ \ (i \neq j).$$

Consequences:

$$\mathrm{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}, \qquad \hat{\sigma}^2 = \frac{e'e}{n-k}.$$

Error Covariance: General Case

$$\mathbb{E}[\varepsilon\varepsilon' \mid X] = \Omega \quad \text{(unknown, p.s.d.).}$$

- If $\mathbb{E}[\varepsilon \mid X] = 0$, OLS remains *unbiased*.
- Variance changes:
$$\mathrm{Var}(\hat{\beta} \mid X) = (X'X)^{-1}X'\,\Omega\,X\,(X'X)^{-1}.$$
- Heteroskedasticity-robust SEs (White):
$$\widehat{\mathrm{Var}}(\hat{\beta}) = (X'X)^{-1}X'\,(ee')\,X\,(X'X)^{-1}.$$

Normality (for Tests)

For exact small-sample t/F tests it is convenient to assume

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n) \quad \text{(given } X\text{)}.$$

This is *not* needed for unbiasedness or BLUE; with large n, asymptotic normality of $\hat{\beta}$ yields approximate t/F inference via the CLT.

Classical Linear Model (CLM) Assumptions

Scalar notation	Matrix notation
1. $\mathbb{E}(\varepsilon_i) = 0$ for each i.	1. $\mathbb{E}(\varepsilon) = 0$.
2. $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$; $\mathbb{E}(\varepsilon_i^2) = \sigma^2$.	2. $\mathbb{E}(\varepsilon\varepsilon') = \sigma^2 I_n$.
3. $X_2, \ldots, X_k$ treated as fixed.	3. X treated as nonstochastic (work conditional on X).
4. No exact linear relation among the X's.	4. $\text{rank}(X) = k < n$ (full column rank).
5. For tests: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.	5. For tests: $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

What Each Assumption Gives Us

Assumption	**Implication**
$\text{rank}(X) = k$	Uniqueness of $\hat{\beta}$.
$\mathbb{E}[\varepsilon \mid X] = 0$	$\mathbb{E}[\hat{\beta} \mid X] = \beta$ (unbiased).
$\mathbb{E}[\varepsilon\varepsilon' \mid X] = \sigma^2 I_n$	BLUE; $\text{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$.
Normality (or CLT)	Exact (or asymptotic) t/F inference.

Gauss–Markov (BLUE)

Under linearity in parameters, full column rank, $\mathbb{E}[\varepsilon \mid X] = 0$, and $\mathbb{E}[\varepsilon\varepsilon' \mid X] = \sigma^2 I_n$:

$$\hat{\beta} = (X'X)^{-1}X'y \quad \text{is Best Linear Unbiased (BLUE).}$$

"Best": minimum variance among all linear unbiased estimators.

Appendix

Matrix Derivative Rules Used

Let a, b be $K \times 1$ vectors and A a symmetric $K \times K$ matrix. We use:

$$\frac{\partial\, a'b}{\partial b} = a, \qquad \frac{\partial\, b'a}{\partial b} = a, \qquad \frac{\partial\, b'Ab}{\partial b} = 2Ab.$$

In our OLS derivation:

$$\frac{\partial\, \hat{\beta}'X'y}{\partial \hat{\beta}} = X'y, \qquad \frac{\partial\, \left(-2\,\hat{\beta}'X'y\right)}{\partial \hat{\beta}} = -2X'y, \qquad \frac{\partial\, \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}} = 2X'X\,\hat{\beta}.$$

These identities produce the gradient $\nabla_{\hat{\beta}}(e'e) = -2X'y + 2X'X\hat{\beta}$.

Linear Independence

Let $\{x_1, \ldots, x_r\}$ be $n \times 1$ column vectors. **Definition:** $\{x_1, \ldots, x_r\}$ is *linearly independent* iff

$$\alpha_1 x_1 + \cdots + \alpha_r x_r = 0 \;\Rightarrow\; \alpha_1 = \cdots = \alpha_r = 0.$$

Equivalent: no vector in the set can be written as a linear combination of the others.

Example. $\begin{bmatrix} 1 & 3 \\ 2 & 6 \\ 0 & 0 \end{bmatrix}$ has dependent columns (second is $3\times$ the first).

Rank of a Matrix

For $A \in \mathbb{R}^{n \times m}$ with columns $a_1, \ldots, a_m$:

$$\operatorname{rank}(A) = \max\{\text{number of linearly independent columns of } A\} = \dim \mathcal{C}(A).$$

Basic facts:

- $\operatorname{rank}(A) \leq \min(n, m)$, and $\operatorname{rank}(A) = \operatorname{rank}(A')$.
- $\operatorname{rank}(A) = m \Rightarrow A$ has *full column rank*.
- If A is $k \times k$ and $\operatorname{rank}(A) = k$, then A is invertible.

Why Rank Matters for OLS

Let $X = [x_1 \cdots x_k] \in \mathbb{R}^{n \times k}$.

- $\operatorname{rank}(X) = k$ (columns independent) $\Rightarrow X'X$ is $k \times k$ and invertible.
- Reason: for any nonzero z, $z'X'Xz = \|Xz\|^2 > 0 \Rightarrow X'X$ is positive definite.
- Then the normal equations have a unique solution:

$$(X'X)\hat{\beta} = X'y \quad \Rightarrow \quad \hat{\beta} = (X'X)^{-1}X'y.$$

- If $\operatorname{rank}(X) < k$: perfect multicollinearity $\Rightarrow X'X$ singular, coefficients not identified.