Linear Models: Geometry and OLS

Statistics · Module 8 — Multivariate Linear Models

Node 8.3

Position in Knowledge Graph

System: Statistics. Structural Domain: Module 8 — Multivariate and Linear Models. Node 8.3 — Linear Models: Geometry and OLS.

The linear model is the most fundamental regression framework in statistics. This node develops OLS as an orthogonal projection, establishes the geometric interpretation, and derives the properties of the hat matrix and residuals. It depends on the covariance structure (Node 8.1) and the MVN (Node 8.2), and enables Gauss-Markov (Node 8.4), inference (Node 8.5), and GLMs (Node 8.6).

Formal Definition

The linear model is

$$Y = X\beta + \varepsilon,$$

where $Y \in \mathbb{R}^n$ is the response, $X \in \mathbb{R}^{n \times p}$ is the design matrix with $\text{rank}(X) = p \leq n$, $\beta \in \mathbb{R}^p$ is the coefficient vector, and $\varepsilon \in \mathbb{R}^n$ satisfies $\mathbb{E}[\varepsilon] = 0$, $\text{Cov}(\varepsilon) = \sigma^2 I_n$.

The **OLS estimator** is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Intuition

The column space $\mathcal{C}(X) = \{X\beta : \beta \in \mathbb{R}^p\}$ is a p-dimensional subspace of $\mathbb{R}^n$. The model says the mean response $\mathbb{E}[Y] = X\beta$ lies in this subspace. OLS finds the point in $\mathcal{C}(X)$ closest to the observed Y in Euclidean distance — this is orthogonal projection. The residuals are the component of Y perpendicular to $\mathcal{C}(X)$, and the fitted values are the projection. This geometric view makes the algebraic properties of OLS transparent.

Structural Role

The linear model separates signal $X\beta \in \mathcal{C}(X)$ from noise $\varepsilon \perp \mathcal{C}(X)$ via orthogonal projection. The hat matrix H is the central object: it encodes both the estimator and the residuals, and its rank determines degrees of freedom. Optimality of OLS under second-moment assumptions is established in Node 8.4. Under normality, OLS is also the MLE, and exact finite-sample inference follows (Node 8.5).

Mathematical Development

Derivation of the normal equations. The OLS criterion is $\min_\beta \|Y - X\beta\|^2$. Expanding:

$$\|Y - X\beta\|^2 = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta.$$

Taking the gradient and setting to zero:

$$\nabla_\beta \|Y - X\beta\|^2 = -2X^\top Y + 2X^\top X\beta = 0.$$

This gives the **normal equations** $X^\top X\hat{\beta} = X^\top Y$. Since $\text{rank}(X) = p$, $X^\top X$ is invertible and $\hat{\beta} = (X^\top X)^{-1}X^\top Y$.

Geometric interpretation. The normal equations $X^\top(Y - X\hat{\beta}) = 0$ state that the residual $Y - X\hat{\beta}$ is orthogonal to every column of X, hence orthogonal to $\mathcal{C}(X)$. This is the defining property of orthogonal projection.

Hat matrix. The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y = HY,$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix.

Properties of H: - Symmetric: $H^\top = H$. - Idempotent: $H^2 = H$ (projecting twice is the same as projecting once). - $\text{rank}(H) = \text{tr}(H) = p$. - $HX = X$ (columns of X are in the range of H). - Eigenvalues of H are 0 or 1, with exactly p ones.

Residuals. The residuals are

$$\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y.$$

The matrix $I_n - H$ is also symmetric and idempotent, with rank $n - p$. Key property: $(I - H)X = 0$, so the residuals are orthogonal to every column of X.

Orthogonal decomposition. The observation vector decomposes as

$$Y = HY + (I - H)Y = \hat{Y} + \hat{\varepsilon},$$

with $\hat{Y} \perp \hat{\varepsilon}$. The Pythagorean theorem gives

$$\|Y\|^2 = \|\hat{Y}\|^2 + \|\hat{\varepsilon}\|^2.$$

After centering (subtracting the mean), this becomes the ANOVA decomposition: TSS = ESS + RSS.

Moments of the OLS estimator.

$$\mathbb{E}[\hat{\beta}] = (X^\top X)^{-1}X^\top \mathbb{E}[Y] = (X^\top X)^{-1}X^\top X\beta = \beta.$$

$$\text{Cov}(\hat{\beta}) = (X^\top X)^{-1}X^\top \text{Cov}(Y)X(X^\top X)^{-1} = \sigma^2(X^\top X)^{-1}.$$

OLS is unbiased for any error distribution with mean zero and any σ^2.

Residual sum of squares.

$$\mathrm{RSS} = \|\hat{\varepsilon}\|^2 = Y^\top (I - H) Y.$$

Since $\mathbb{E}[Y^\top (I-H)Y] = \mathrm{tr}((I-H)\sigma^2 I) + \beta^\top X^\top (I-H) X\beta = \sigma^2(n-p)$ (using $(I-H)X = 0$), we get

$$\mathbb{E}[\mathrm{RSS}] = \sigma^2(n-p),$$

so $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$ is unbiased for σ^2.

Leverage scores. The diagonal elements $h_{ii} = H_{ii}$ are the leverage scores. They satisfy $0 \le h_{ii} \le 1$ and $\sum_i h_{ii} = p$. The leverage h_{ii} measures the influence of the i-th observation on its own fitted value: $\hat{Y}_i = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j$. High-leverage points have h_{ii} close to 1 and dominate their own fit.

Residual properties. Under homoskedasticity, $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(I-H)$, so $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1-h_{ii})$. The residuals are heteroskedastic even when the errors are not. Standardized residuals $r_i = \hat{\varepsilon}_i / (\hat{\sigma}\sqrt{1-h_{ii}})$ correct for this.

Connection to the Moore-Penrose pseudoinverse. When X has full column rank, $(X^\top X)^{-1} X^\top = X^+$ is the Moore-Penrose pseudoinverse. The OLS solution is $\hat{\beta} = X^+ Y$, the minimum-norm least-squares solution.

Worked Examples

Example 1 (Simple linear regression). Let $X = (\mathbf{1}, x) \in \mathbb{R}^{n \times 2}$ with $x = (x_1, \ldots, x_n)^\top$. Then

$$X^\top X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad X^\top Y = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}.$$

Solving gives $\hat{\beta}_1 = \sum (x_i - \bar{x})(Y_i - \bar{Y}) / \sum (x_i - \bar{x})^2$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$. The hat matrix has $h_{ii} = 1/n + (x_i - \bar{x})^2 / \sum (x_j - \bar{x})^2$, showing that leverage increases with distance from $\bar{x}$.

Example 2 (One-way ANOVA). For k groups with n_j observations each, X is a block-diagonal indicator matrix. The hat matrix projects onto group means: $\hat{Y}_{ij} = \bar{Y}_{j\cdot}$. The residual is $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{j\cdot}$. RSS $= \sum_j \sum_i (Y_{ij} - \bar{Y}_{j\cdot})^2$ with $n-k$ degrees of freedom, and ESS $= \sum_j n_j (\bar{Y}_{j\cdot} - \bar{Y})^2$ with $k-1$ degrees of freedom.

Cross-Links

AI. Linear regression is empirical risk minimization under squared loss. The hat matrix appears in leave-one-out cross-validation via the shortcut $\mathrm{CV} = n^{-1} \sum (\hat{\varepsilon}_i / (1 - h_{ii}))^2$, avoiding n refits. Ridge regression replaces H with $X(X^\top X + \lambda I)^{-1} X^\top$, shrinking the projection.

Economics. The linear model is the workhorse of empirical economics. OLS with heteroskedasticity-robust standard errors (Huber-White) replaces $\sigma^2 (X^\top X)^{-1}$ with the sandwich covariance (Node 7.5). Instrumental variables estimation projects onto a different column space.

Linear Algebra. OLS is the orthogonal projection theorem in $\mathbb{R}^n$ with the standard inner product. The normal equations are the first-order optimality conditions for a convex quadratic. The hat matrix is a rank-p orthogonal projector, decomposing $\mathbb{R}^n = \mathcal{C}(X) \oplus \mathcal{C}(X)^\perp$.

Quantum Mechanics. In quantum state tomography, linear inversion estimators reconstruct the density matrix by projecting measurement outcomes onto the space of valid states, the same projection structure as OLS but on the cone of positive semidefinite matrices.

Structural Compression

Invariant. $\hat{Y} = HY$ is the orthogonal projection of Y onto $\mathcal{C}(X)$; the OLS estimator is the unique solution to the normal equations $X^\top X\hat{\beta} = X^\top Y$.

Mechanism. Minimizing $\|Y - X\beta\|^2$ over β is equivalent to finding the closest point in $\mathcal{C}(X)$ to Y; the projection theorem yields the hat matrix as the unique orthogonal projector. The orthogonal decomposition $Y = \hat{Y} + \hat{\varepsilon}$ separates signal from noise.

Cross-System Echo. OLS is the Moore-Penrose pseudoinverse solution. In numerical linear algebra, the QR decomposition provides a numerically stable computation: $X = QR$ gives $\hat{\beta} = R^{-1}Q^\top Y$ and $H = QQ^\top$. In AI, the leverage scores h_{ii} determine the effective number of parameters and the generalization error via the optimism.

Exercises

1. Derive the OLS estimator by completing the square in $\|Y - X\beta\|^2$ rather than by differentiation.
2. Show that H is the unique symmetric idempotent matrix with $\mathcal{C}(H) = \mathcal{C}(X)$.
3. Prove that $\sum_{i=1}^{n} h_{ii} = p$ using $\mathrm{tr}(H) = \mathrm{tr}(X(X^\top X)^{-1}X^\top)$ and the cyclic property of trace.
4. For simple linear regression with equally spaced $x_i = i$, compute the leverage scores and identify the points with maximum leverage.
5. Show that adding a predictor to the model (increasing p by 1) always decreases RSS, and relate this to the geometry of the projection.
6. Prove the leave-one-out formula: the i-th cross-validation residual is $\hat{\varepsilon}_i^{(-i)} = \hat{\varepsilon}_i/(1 - h_{ii})$ using the Sherman-Morrison-Woodbury formula.
7. Argue, structurally, why the hat matrix is the fundamental object in the linear model — more fundamental than $\hat{\beta}$ — by showing that all inferential quantities (fitted values, residuals, RSS, leverage, degrees of freedom) are determined by H alone.

Statistics · Module 8 — Multivariate Linear Models · Node 8.3
Concepts: vector-space, orthogonality, linear-operators, matrix-representation
Prerequisites: 8.1, 8.2
Enables: 8.4, 8.5, 8.6

hbar.university — own.your.cognition