Linear Regression - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Linear regression models a real-valued response as an affine function of the features. The estimator is obtained by minimizing the (empirical) squared loss, yielding the *ordinary least squares* (OLS) solution in closed form. We present the optimization and geometric derivations (normal equations and projection), state conditions for existence/uniqueness, give weighted/regularized variants, and summarize finite-sample statistical properties (Gauss–Markov). All variables and dimensions follow the notation in this section.

1.2 Data and Notation

Given training data

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \qquad \mathbf{x}_i \in \mathbb{R}^d \ \ (\text{column}, d \times 1), \quad y_i \in \mathbb{R} \ (\text{scalar}).$$

To include an intercept, augment each feature vector by a leading 1:

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix} \in \mathbb{R}^p, \quad p = d + 1, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_{1:d} \end{bmatrix} \in \mathbb{R}^p.$$

Stack the predictors and responses into the *design matrix* and response vector:

$$\mathbf{X} = \begin{bmatrix} \tilde{\mathbf{x}}_1^\top \\ \vdots \\ \tilde{\mathbf{x}}_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

Dimensions: $n, d, p \in \mathbb{N}$; $\mathbf{X} \in \mathbb{R}^{n \times p}$; $\mathbf{y} \in \mathbb{R}^n$; $\boldsymbol{\beta} \in \mathbb{R}^p$.

1.3 Model Formulation and OLS Objective

The linear model assumes

$$\widehat{y}_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_{1:d} \;=\; \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}, \qquad \widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}.$$

Ordinary least squares minimizes the empirical squared error

$$J(\boldsymbol{\beta}) \;=\; \frac{1}{2}\,\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \;=\; \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \tilde{\mathbf{x}}_i^\top \boldsymbol{\beta}\right)^2.$$

(The factor $\frac{1}{2}$ is for algebraic convenience; it does not change the minimizer.)

1.4 Calculus Derivation: Normal Equations and Solution

Expand J and differentiate:

$$J(\boldsymbol{\beta}) = \tfrac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \tfrac{1}{2}\mathbf{y}^\top\mathbf{y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \tfrac{1}{2}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

Gradient and Hessian:

$$\nabla J(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^\top\mathbf{y}, \qquad \nabla^2 J(\boldsymbol{\beta}) = \mathbf{X}^\top\mathbf{X} \;\succeq\; \mathbf{0}.$$

Setting the gradient to zero gives the *normal equations*

$$\boxed{\mathbf{X}^\top\mathbf{X}\,\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}}\,.$$

Existence/uniqueness. If $\operatorname{rank}(\mathbf{X}) = p$ (full column rank), then $\mathbf{X}^\top\mathbf{X} \in \mathbb{R}^{p\times p}$ is positive definite and invertible, and

$$\boxed{\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}}\,.$$

If $\operatorname{rank}(\mathbf{X}) < p$, the set of minimizers is affine; the minimum-*norm* solution is

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^{+}\mathbf{y},$$

where $\mathbf{X}^{+}$ is the Moore–Penrose pseudoinverse (e.g., via SVD).

1.5 Geometric Derivation: Orthogonal Projection

Let $\mathcal{C}(\mathbf{X}) \subseteq \mathbb{R}^n$ be the column space of $\mathbf{X}$. Any fitted vector $\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ lies in $\mathcal{C}(\mathbf{X})$. OLS chooses the point $\widehat{\mathbf{y}} \in \mathcal{C}(\mathbf{X})$ closest (in ℓ_2) to $\mathbf{y}$, i.e., the *orthogonal projection* of $\mathbf{y}$ onto $\mathcal{C}(\mathbf{X})$.

Normal equations as orthogonality. Let $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and residual $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$. Then $\widehat{\boldsymbol{\beta}}$ is optimal iff

$$\mathbf{X}^\top\mathbf{r} = \mathbf{0} \qquad \text{(residual orthogonal to all columns of } \mathbf{X}\text{)},$$

which is equivalent to the normal equations.

Projection (hat) matrix. If $\operatorname{rank}(\mathbf{X}) = p$,

$$\mathbf{P}_X \;\triangleq\; \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \quad\Rightarrow\quad \widehat{\mathbf{y}} = \mathbf{P}_X\,\mathbf{y}, \;\; \mathbf{r} = (\mathbf{I} - \mathbf{P}_X)\mathbf{y}.$$

Properties: $\mathbf{P}_X$ is symmetric and idempotent ($\mathbf{P}_X^\top = \mathbf{P}_X$, $\mathbf{P}_X^2 = \mathbf{P}_X$); $\operatorname{rank}(\mathbf{P}_X) = p$; $(\mathbf{I} - \mathbf{P}_X)$ projects onto $\mathcal{C}(\mathbf{X})^\perp$.

1.6 Statistical Model and Properties (Gauss–Markov)

Assume the classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n.$$

Then the OLS estimator is unbiased and has covariance

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}^\star, \qquad \mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

An unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\|\mathbf{r}\|_2^2}{n-p} = \frac{1}{n-p}\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}).$$

Gauss–Markov (BLUE) theorem. Among all *linear unbiased* estimators $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$ with $\mathbf{A}\mathbf{X} = \mathbf{I}_p$, $\widehat{\boldsymbol{\beta}}$ has the smallest covariance (Loewner order):

$$\mathrm{Var}(\tilde{\boldsymbol{\beta}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}}) \ \succeq\ \mathbf{0}.$$

Proof. Any linear unbiased $\tilde{\boldsymbol{\beta}}$ can be written as $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + \mathbf{D}\mathbf{y}$ with $\mathbf{D}\mathbf{X} = \mathbf{0}$. Then $\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = \mathrm{Var}(\widehat{\boldsymbol{\beta}}) + \sigma^2\mathbf{D}\mathbf{D}^\top \ \succeq\ \mathrm{Var}(\widehat{\boldsymbol{\beta}})$. ■

ANOVA / R^2 (with intercept). Let $\bar{y} = \frac{1}{n}\sum_i y_i$. Total sum of squares $\mathrm{SST} = \sum_i (y_i - \bar{y})^2$ decomposes as $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$ with $\mathrm{SSR} = \sum_i(\widehat{y}_i - \bar{y})^2$, $\mathrm{SSE} = \sum_i(y_i - \widehat{y}_i)^2 = \|\mathbf{r}\|_2^2$, and $R^2 = \mathrm{SSR}/\mathrm{SST} = 1 - \mathrm{SSE}/\mathrm{SST}$.

1.7 Variants: Weighted and Regularized Least Squares

Weighted least squares (known heteroscedasticity). Given positive weights $\{w_i\}$ (or diagonal $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$), minimize $\frac{1}{2}\|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2$. The solution satisfies $(\mathbf{X}^\top\mathbf{W}\mathbf{X})\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{W}\mathbf{y}$, i.e.,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}.$$

Ridge (Tikhonov) regularization. Minimize $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2$ (with β_0 typically unpenalized by excluding its column from the penalty). Normal equations become

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)\,\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Ridge is well-defined even when $\mathrm{rank}(\mathbf{X}) < p$.

1.8 Computation and Algorithms

Direct solvers. Avoid explicit inverses; solve normal equations by Cholesky if $\mathbf{X}^\top\mathbf{X}$ is well-conditioned, or use a QR decomposition $\mathbf{X} = \mathbf{Q}\mathbf{R}$ with $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$: $\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{y}$. For rank-deficient problems or better numerical stability, use SVD: $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \Rightarrow \widehat{\boldsymbol{\beta}} = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top\mathbf{y}$.

First-order (large-scale). Gradient descent on J: $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \eta\,\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}^{(t)} - \mathbf{y})$, with $0 < \eta < 2/L$ where $L = \|\mathbf{X}^\top\mathbf{X}\|_{\mathrm{op}}$. Variants: stochastic/mini-batch, conjugate gradients (on normal equations).

1.9 Intercept Handling and Centering

If the first column of $\mathbf{X}$ is all ones, the OLS intercept equals the residual mean constraint $\sum_i r_i = 0$. Alternatively, center columns of $\mathbf{X}$ and $\mathbf{y}$: then $\widehat{\beta}_0 = \bar{y}$ and the slope coefficients are obtained by regressing the centered $\mathbf{y}$ on centered $\mathbf{X}$.

1.10 Algorithm (Ordinary Least Squares)

1. **Input:** data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, with optional intercept column added to $\mathbf{X}$.
2. **Form** $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \in \mathbb{R}^n$.
3. **Solve** the normal equations for $\widehat{\boldsymbol{\beta}}$ via QR/SVD (or Cholesky on $\mathbf{X}^\top\mathbf{X}$).
4. **Predict** $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$; compute residuals $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$.
5. **Diagnostics:** $\widehat{\sigma}^2 = \|\mathbf{r}\|_2^2/(n-p)$, leverage $h_{ii} = (\mathbf{P}_X)_{ii}$, studentized residuals, R^2 (if intercept present).

1.11 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: ith feature vector (dimension $d \times 1$).
- $\tilde{\mathbf{x}}_i \in \mathbb{R}^p$: augmented feature with intercept ($p = d + 1$).
- $y_i \in \mathbb{R}$: response for sample i (scalar). $\quad \mathbf{y} \in \mathbb{R}^n$.
- $\mathbf{X} \in \mathbb{R}^{n\times p}$: design matrix (rows $\tilde{\mathbf{x}}_i^\top$).
- $\boldsymbol{\beta} \in \mathbb{R}^p$: parameter vector; β_0 intercept, $\boldsymbol{\beta}_{1:d}$ slopes.
- $\widehat{\boldsymbol{\beta}}$: OLS solution; $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ if $\mathrm{rank}(\mathbf{X}) = p$; $\mathbf{X}^+\mathbf{y}$ otherwise.
- $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$: fitted values; $\mathbf{r} = \mathbf{y} - \widehat{\mathbf{y}}$ residuals.
- $\mathbf{P}_X = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$: projection (hat) matrix ($n \times n$).
- σ^2: noise variance; $\widehat{\sigma}^2 = \|\mathbf{r}\|_2^2/(n-p)$ unbiased estimator (with intercept).
- For WLS: $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$, solution $(\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}$.
- For ridge: penalty $\lambda \geq 0$, solution $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y}$.

1.12 Summary

From first principles, linear regression selects $\widehat{\boldsymbol{\beta}}$ to minimize squared loss, leading to the normal equations $\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^\top\mathbf{y}$. Geometrically, OLS projects $\mathbf{y}$ onto the column space of $\mathbf{X}$ via the hat matrix, with residuals orthogonal to all regressors. Under homoscedastic uncorrelated noise, OLS is unbiased and achieves minimum variance among linear unbiased estimators (Gauss–Markov). Extensions include WLS for known heteroscedasticity and ridge for regularization; numerically stable computation relies on QR/SVD rather than explicit matrix inversion.