Chapter 1

Linear Regression

Ordinary Least Squares

Theorem 1.0.1 (Uniqueness of the Least Squares Solution)**.** *Let $\Phi \in \mathbb{R}^{N\times M}$ denote the design matrix and $t \in \mathbb{R}^N$ the target vector. Consider the least squares cost function*

$$E(w) = \frac{1}{2}\|t - \Phi w\|^2.$$

Then:

(i) The function $E(w)$ is convex in w.

(ii) If $\Phi^\top\Phi$ is invertible (i.e., $\mathrm{rank}(\Phi) = M$), then $E(w)$ is strictly convex *and admits a unique minimizer*

$$w^* = (\Phi^\top\Phi)^{-1}\Phi^\top t.$$

(iii) If $\Phi^\top\Phi$ is singular, the minimizer is not unique; all minimizers are of the form

$$w = w_0 + v, \qquad v \in \mathrm{Null}(\Phi),$$

where w_0 is any particular solution to the normal equations $\Phi^\top\Phi w = \Phi^\top t$.

Proof. We begin by expanding the objective:

$$E(w) = \frac{1}{2}(t - \Phi w)^\top(t - \Phi w) = \frac{1}{2}(t^\top t - 2t^\top\Phi w + w^\top\Phi^\top\Phi w).$$

(1) Gradient and Stationary Point: The gradient of $E(w)$ with respect to w is

$$\nabla_w E(w) = -\Phi^\top t + \Phi^\top\Phi w.$$

Setting $\nabla_w E(w) = 0$ yields the *normal equations*

$$\Phi^\top\Phi w = \Phi^\top t. \tag{1}$$

(2) Hessian and Convexity: The Hessian of $E(w)$ is

$$H = \nabla_w^2 E(w) = \Phi^\top\Phi.$$

For any nonzero vector $z \in \mathbb{R}^M$,

$$z^\top H z = z^\top\Phi^\top\Phi z = \|\Phi z\|^2 \geq 0,$$

hence H is positive semidefinite, implying $E(w)$ is convex.
If Φ has full column rank ($\mathrm{rank}(\Phi) = M$), then $\Phi^\top\Phi$ is positive definite, and

$$z^\top H z = 0 \quad \Leftrightarrow \quad z = 0,$$

so $E(w)$ is strictly convex. A strictly convex function has a unique minimizer, obtained by solving (1):

$$w^* = (\Phi^\top\Phi)^{-1}\Phi^\top t.$$

(3) Non-uniqueness for Rank-Deficient Φ: If $\Phi^\top\Phi$ is singular, there exist nonzero vectors v such that $\Phi v = 0$. For any particular solution w_0 satisfying (1), we have

$$\Phi^\top\Phi(w_0 + v) = \Phi^\top\Phi w_0 + \Phi^\top\Phi v = \Phi^\top t,$$

since $\Phi v = 0$. Thus, every vector $w = w_0 + v$, with $v \in \mathrm{Null}(\Phi)$, minimizes $E(w)$. The minimal-norm solution among them is given by the Moore–Penrose pseudoinverse:

$$w^* = \Phi^+ t.$$

(4) Conclusion: The cost $E(w)$ is convex for all Φ, and strictly convex (hence uniquely minimized) iff $\Phi^\top\Phi$ is invertible. □

Theorem 1.0.2 (Unbiasedness of the OLS Estimator)**.** *Assume the linear regression model*

$$t = \Phi w + \varepsilon,$$

where $\Phi \in \mathbb{R}^{N\times M}$ is the design matrix, $w \in \mathbb{R}^M$ the true parameter vector, and the noise satisfies $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. Assume further that $\Phi^\top\Phi$ is invertible. Then the ordinary least squares estimator

$$\hat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top t$$

is an unbiased estimator of w, i.e.

$$\mathbb{E}[\hat{w}] = w.$$

Proof. By the model,

$$t = \Phi w + \varepsilon.$$

Substitute into the estimator:

$$\hat{w} = (\Phi^\top \Phi)^{-1} \Phi^\top t = (\Phi^\top \Phi)^{-1} \Phi^\top (\Phi w + \varepsilon).$$

Distribute terms:

$$\hat{w} = (\Phi^\top \Phi)^{-1} \Phi^\top \Phi\, w \;+\; (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon.$$

Since $(\Phi^\top \Phi)^{-1} \Phi^\top \Phi = I_M$, this simplifies to

$$\hat{w} = w + (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon.$$

Take expectation using linearity and $\mathbb{E}[\varepsilon] = 0$:

$$\mathbb{E}[\hat{w}] = \mathbb{E}\big[w + (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon\big] = w + (\Phi^\top \Phi)^{-1} \Phi^\top \mathbb{E}[\varepsilon] = w + (\Phi^\top \Phi)^{-1} \Phi^\top 0 = w.$$

Thus $\hat{w}$ is unbiased. □

Corollary 1.0.3. *Under the same assumptions,*

$$\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

Proof. From $\hat{w} = w + (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon$ and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$,

$$\mathrm{Cov}(\hat{w}) = (\Phi^\top \Phi)^{-1} \Phi^\top \,\mathrm{Cov}(\varepsilon)\, \Phi (\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1} \Phi^\top \Phi (\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

□

Theorem 1.0.4 (Covariance of the OLS Estimator)**.** *Under the linear regression model*

$$t = \Phi w + \varepsilon, \qquad \mathbb{E}[\varepsilon] = 0, \quad \mathrm{Cov}(\varepsilon) = \sigma^2 I,$$

with $\Phi \in \mathbb{R}^{N \times M}$ of full column rank, the ordinary least squares estimator

$$\hat{w} = (\Phi^\top \Phi)^{-1} \Phi^\top t$$

has covariance matrix

$$\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}.$$

Proof. From the model $t = \Phi w + \varepsilon$,

$$\hat{w} = (\Phi^\top \Phi)^{-1} \Phi^\top t = (\Phi^\top \Phi)^{-1} \Phi^\top (\Phi w + \varepsilon) = w + (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon.$$

Subtract the expectation $\mathbb{E}[\hat{w}] = w$ to get the deviation:

$$\hat{w} - \mathbb{E}[\hat{w}] = (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon.$$

Now compute the covariance:

$$\begin{aligned} \mathrm{Cov}(\hat{w}) &= \mathbb{E}\big[(\hat{w} - \mathbb{E}[\hat{w}])(\hat{w} - \mathbb{E}[\hat{w}])^\top\big] \\ &= \mathbb{E}\big[(\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon \varepsilon^\top \Phi (\Phi^\top \Phi)^{-1}\big]. \end{aligned}$$

Using $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ and the linearity of expectation:

$$\mathrm{Cov}(\hat{w}) = (\Phi^\top \Phi)^{-1} \Phi^\top (\sigma^2 I) \Phi (\Phi^\top \Phi)^{-1} = \sigma^2 (\Phi^\top \Phi)^{-1} \Phi^\top \Phi (\Phi^\top \Phi)^{-1}.$$

Simplifying:

$$\boxed{\mathrm{Cov}(\hat{w}) = \sigma^2 (\Phi^\top \Phi)^{-1}}.$$

□

Theorem 1.0.5 (Gauss–Markov Theorem)**.** *Consider the linear model*

$$t = \Phi w + \varepsilon,$$

with $\Phi \in \mathbb{R}^{N \times M}$ of full column rank, $\mathbb{E}[\varepsilon] = 0$, and $\mathrm{Cov}(\varepsilon) = \sigma^2 I$. Let $\hat{w}_{\mathrm{OLS}} = (\Phi^\top \Phi)^{-1} \Phi^\top t$ denote the ordinary least squares estimator. Then $\hat{w}_{\mathrm{OLS}}$ is the Best Linear Unbiased Estimator (BLUE)*: for any other linear unbiased estimator of the form $\tilde{w} = Ct$ (with constant matrix $C \in \mathbb{R}^{M \times N}$ such that $\mathbb{E}[\tilde{w}] = w$), we have*

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}_{\mathrm{OLS}}) \;\succeq\; 0,$$

i.e. the matrix difference is positive semidefinite. Equivalently, every componentwise variance of $\tilde{w}$ is at least that of $\hat{w}_{\mathrm{OLS}}$.

Proof. Let $\tilde{w}$ be any linear estimator of the form $\tilde{w} = Ct$ for a fixed matrix $C \in \mathbb{R}^{M \times N}$. The unbiasedness condition $\mathbb{E}[\tilde{w}] = w$ requires

$$\mathbb{E}[Ct] = C\mathbb{E}[t] = C\Phi w = w \quad \text{for all } w,$$

hence

$$C\Phi = I_M. \tag{1}$$

Write the OLS estimator as

$$\hat{w} \equiv \hat{w}_{\mathrm{OLS}} = (\Phi^\top \Phi)^{-1} \Phi^\top t.$$

Define the matrix difference

$$A \;:=\; C - (\Phi^\top \Phi)^{-1} \Phi^\top.$$

Using (1) and the identity $\big((\Phi^\top \Phi)^{-1} \Phi^\top\big)\Phi = I_M$, we obtain

$$A\Phi = C\Phi - (\Phi^\top \Phi)^{-1} \Phi^\top \Phi = I_M - I_M = 0.$$

Thus

$$A\Phi = 0 \qquad \Longrightarrow \qquad A\Phi w = 0 \quad \text{for all } w.$$

Now express $\tilde{w}$ in terms of $\hat{w}$ and A:

$$\tilde{w} = Ct = \big((\Phi^\top \Phi)^{-1} \Phi^\top + A\big) t = \hat{w} + At.$$

Subtracting expectations (and using $\mathbb{E}[\hat{w}] = \mathbb{E}[\tilde{w}] = w$) gives the zero-mean deviations

$$\tilde{w} - w = (\hat{w} - w) + A\varepsilon,$$

since $t = \Phi w + \varepsilon$ and $A\Phi w = 0$.
Compute the covariance matrices. Using $\mathrm{Cov}(\varepsilon) = \sigma^2 I$ and independence of deterministic matrices from ε,

$$\begin{aligned} \mathrm{Cov}(\tilde{w}) &= \mathbb{E}\big[(\tilde{w} - w)(\tilde{w} - w)^\top\big] \\ &= \mathbb{E}\big[(\hat{w} - w + A\varepsilon)(\hat{w} - w + A\varepsilon)^\top\big] \\ &= \mathrm{Cov}(\hat{w}) + A\,\mathbb{E}[\varepsilon\varepsilon^\top]\,A^\top \;+\; \mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big] A^\top \;+\; A\,\mathbb{E}\big[\varepsilon (\hat{w} - w)^\top\big]. \end{aligned}$$

□

But $\hat{w} - w = (\Phi^\top \Phi)^{-1}\Phi^\top \varepsilon$ is linear in ε, so

$$\mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big] = (\Phi^\top\Phi)^{-1}\Phi^\top \mathbb{E}[\varepsilon\varepsilon^\top] = (\Phi^\top\Phi)^{-1}\Phi^\top(\sigma^2 I) = \sigma^2(\Phi^\top\Phi)^{-1}\Phi^\top.$$

Since $A\Phi = 0$, we have

$$\mathbb{E}\big[(\hat{w} - w)\varepsilon^\top\big]A^\top = \sigma^2(\Phi^\top\Phi)^{-1}\Phi^\top A^\top = \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top A^\top) = \sigma^2(\Phi^\top\Phi)^{-1}(A\Phi)^\top = 0.$$

Similarly the other cross term $A\,\mathbb{E}[\varepsilon(\hat{w} - w)^\top]$ vanishes. Thus the covariance simplifies to

$$\mathrm{Cov}(\tilde{w}) = \mathrm{Cov}(\hat{w}) + A\,\mathbb{E}[\varepsilon\varepsilon^\top]\,A^\top = \mathrm{Cov}(\hat{w}) + \sigma^2 AA^\top.$$

Therefore

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}) = \sigma^2 AA^\top.$$

But $\sigma^2 AA^\top$ is positive semidefinite (for any $\sigma^2 \geq 0$ and any matrix A), so

$$\mathrm{Cov}(\tilde{w}) - \mathrm{Cov}(\hat{w}) \succeq 0,$$

which proves that $\hat{w}$ has the smallest covariance matrix among all linear unbiased estimators. This completes the proof. □

Theorem 1.0.6 (Orthogonality of Residuals). *Let $\Phi \in \mathbb{R}^{N\times M}$ be the design matrix and $t \in \mathbb{R}^N$ the observed targets. Let $\hat{w}$ be any solution of the normal equations*

$$\Phi^\top\Phi\,\hat{w} \;=\; \Phi^\top t.$$

Define the residual vector $r := t - \Phi\hat{w}$. Then

$$\Phi^\top r \;=\; 0,$$

i.e. r is orthogonal to every column of Φ (equivalently r is orthogonal to $\mathrm{col}(\Phi)$).

Proof. Starting from the normal equations,

$$\Phi^\top\Phi\,\hat{w} \;=\; \Phi^\top t.$$

Rearrange terms to move $\Phi^\top\Phi\hat{w}$ to the right-hand side:

$$\Phi^\top t - \Phi^\top\Phi\,\hat{w} \;=\; 0.$$

Factor $\Phi^\top$:

$$\Phi^\top(t - \Phi\hat{w}) \;=\; 0.$$

But $t - \Phi\hat{w}$ is exactly the residual vector r, hence

$$\Phi^\top r \;=\; 0.$$

This shows each column of Φ has zero inner product with r, i.e. $r \perp \mathrm{col}(\Phi)$. □

Corollary 1.0.7 (Hat Matrix and Residual Projection). *If Φ has full column rank and $\hat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top t$, define the hat (projection) matrix*

$$P := \Phi(\Phi^\top\Phi)^{-1}\Phi^\top.$$

Then the fitted values are $\hat{t} = Pt$ and the residual satisfies

$$r = (I - P)t,$$

with $P^2 = P$ and $P^\top = P$. Consequently $(I - P)$ is the orthogonal projector onto $\mathrm{col}(\Phi)^\perp$, and r is the orthogonal projection of t onto that complement.

Proof. Using $\hat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top t$ gives $\hat{t} = \Phi\hat{w} = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top t = Pt$, so $r = t - \hat{t} = (I - P)t$. The identities $P^2 = P$ and $P^\top = P$ follow from straightforward algebra:

$$P^2 = \Phi(\Phi^\top\Phi)^{-1}\underbrace{\Phi^\top\Phi}_{=}(\Phi^\top\Phi)^{-1}\Phi^\top = P, \qquad P^\top = \big(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top\big)^\top = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top = P.$$

Thus P is an orthogonal projector onto $\mathrm{col}(\Phi)$ and $(I - P)$ projects orthogonally onto its complement, so r lies in $\mathrm{col}(\Phi)^\perp$. □

Bayesian Linear Regression: Prior on w and Predictive Distribution

Bayesian Formulation

In Bayesian linear regression we treat the parameter vector w as a random variable and place a prior distribution on it. The generative model is:

$$t = \Phi w + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \beta^{-1} I_N),$$

where β is the noise precision.

Prior Distribution on w

We choose a zero-mean isotropic Gaussian prior:

$$p(w) = \mathcal{N}(w \mid 0, \alpha^{-1} I_M),$$

where α is the prior precision. This encodes the belief that large weights are unlikely (acts as a regularizer).

Likelihood

Conditioned on w, the likelihood of the data is:

$$p(t \mid \Phi, w, \beta) = \mathcal{N}(t \mid \Phi w, \beta^{-1} I_N).$$

Posterior Distribution of w

By Bayes' theorem:

$$p(w \mid t, \Phi) \propto p(t \mid \Phi, w, \beta)\, p(w).$$

Because both prior and likelihood are Gaussian, the posterior is also Gaussian:

$$p(w \mid t, \Phi) = \mathcal{N}(w \mid m_N, S_N),$$

with posterior precision and covariance given by:

$$S_N^{-1} = \alpha I_M + \beta\Phi^\top\Phi, \qquad S_N = (\alpha I_M + \beta\Phi^\top\Phi)^{-1},$$

and the posterior mean:

$$m_N = \beta S_N \Phi^\top t.$$

Interpretation

- m_N is the Bayes estimate of w (posterior mean).
- S_N quantifies uncertainty in the weight estimates.
- As $\alpha \to 0$ (weak prior),

$$m_N \to (\Phi^\top\Phi)^{-1}\Phi^\top t,$$

recovering the ordinary least squares solution.

Predictive Distribution

For a new input x_* with feature vector $\phi_* = \phi(x_*)$, the predictive distribution integrates over the posterior uncertainty in w:

$$p(t_* \mid x_*, t, \Phi) = \int p(t_* \mid x_*, w, \beta)\, p(w \mid t, \Phi)\, dw.$$

The integrand is a product of two Gaussians, so the predictive distribution is Gaussian:

$$p(t_* \mid x_*, t, \Phi) = \mathcal{N}\big(t_* \mid m_N^\top \phi_*,\ \beta^{-1} + \phi_*^\top S_N \phi_*\big).$$

Predictive Mean and Variance

Predictive Mean:

$$\mathbb{E}[t_* \mid x_*, t, \Phi] = m_N^\top \phi_*.$$

Predictive Variance:

$$\mathrm{Var}(t_* \mid x_*, t, \Phi) = \underbrace{\beta^{-1}}_{\text{noise variance}} + \underbrace{\phi_*^\top S_N \phi_*}_{\text{model uncertainty}}.$$

Thus the predictive variance decomposes into:

- aleatoric noise (irreducible), and
- epistemic uncertainty (reduced with more data).

Likelihood Derivation (Gaussian Noise) and MLEs

1. Single-observation likelihood

Assume the data generation model for a single observation:

$$t_n = w^\top \phi(x_n) + \varepsilon_n, \qquad \varepsilon_n \sim \mathcal{N}(0, \beta^{-1}).$$

Then the conditional density (likelihood) for t_n given w is

$$p(t_n \mid x_n, w, \beta) = \mathcal{N}\big(t_n \mid w^\top \phi(x_n),\ \beta^{-1}\big) = \sqrt{\frac{\beta}{2\pi}}\, \exp\Big(-\tfrac{\beta}{2}\big(t_n - w^\top \phi(x_n)\big)^2\Big).$$

2. Joint likelihood for the dataset

Assuming i.i.d. noise, the joint likelihood for all N observations is the product

$$p(t \mid \Phi, w, \beta) = \prod_{n=1}^{N} p(t_n \mid x_n, w, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\Big(-\tfrac{\beta}{2}\sum_{n=1}^{N}(t_n - w^\top \phi(x_n))^2\Big).$$

Using matrix notation with $\Phi \in \mathbb{R}^{N\times M}$ and $t \in \mathbb{R}^N$:

$$p(t \mid \Phi, w, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\Big(-\tfrac{\beta}{2}\|t - \Phi w\|^2\Big).$$

3. Log-likelihood

The log-likelihood (more convenient for optimization) is

$$\ell(w, \beta) := \log p(t \mid \Phi, w, \beta) = \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\|t - \Phi w\|^2.$$

Dropping constants independent of the parameters when optimizing:

$$\ell(w, \beta) = \frac{N}{2}\log\beta - \frac{\beta}{2}\|t - \Phi w\|^2 + \text{const}.$$

4. MLE for w (given β)

Take gradient of the log-likelihood w.r.t. w:

$$\nabla_w \ell(w, \beta) = -\frac{\beta}{2}\cdot 2\,(-\Phi^\top)(t - \Phi w) = \beta\Phi^\top(t - \Phi w).$$

Set to zero for critical point:

$$\Phi^\top(t - \Phi w) = 0 \quad \Rightarrow \quad \Phi^\top\Phi\, w = \Phi^\top t.$$

If $\Phi^\top\Phi$ is invertible, the MLE of w is

$$\boxed{\hat{w}_{\mathrm{MLE}} = (\Phi^\top\Phi)^{-1}\Phi^\top t}$$

which is the ordinary least squares solution. Thus MLE = least squares under Gaussian noise.

5. MLE for noise precision β (given w)

Differentiate ℓ w.r.t. β:

$$\frac{\partial \ell}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2}\|t - \Phi w\|^2.$$

Set equal to zero:

$$\frac{N}{2\beta} = \frac{1}{2}\|t - \Phi w\|^2 \quad \Rightarrow \quad \hat{\beta}_{\mathrm{MLE}} = \frac{N}{\|t - \Phi w\|^2}.$$

If we substitute $w = \hat{w}_{\mathrm{MLE}}$ we get the MLE for β:

$$\boxed{\hat{\beta}_{\mathrm{MLE}} = \frac{N}{\|t - \Phi\hat{w}_{\mathrm{MLE}}\|^2}}.$$

Equivalently, the MLE for noise variance $\sigma^2 = \beta^{-1}$ is

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{N}\|t - \Phi\hat{w}_{\mathrm{MLE}}\|^2.$$

(For an unbiased estimator of σ^2 divide by $N - M$ instead of N.)

6. Negative log-likelihood and connection to MAP

The negative log-likelihood (up to additive constant) is

$$-\ell(w, \beta) \propto \frac{\beta}{2}\|t - \Phi w\|^2 - \frac{N}{2}\log\beta.$$

When combining with a Gaussian prior $p(w) \propto \exp\big(-\frac{\alpha}{2}\|w\|^2\big)$, the negative log-posterior (up to constants) becomes

$$-\log p(w \mid t) \propto \frac{\beta}{2}\|t - \Phi w\|^2 + \frac{\alpha}{2}\|w\|^2,$$

whose minimizer yields the MAP estimator. Dividing through by β and setting $\lambda = \alpha/\beta$ gives the familiar ridge form:

$$\hat{w}_{\mathrm{MAP}} = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top t.$$

Derivation of the Posterior with a Gaussian Prior (Completing the Square)

Assume the Gaussian likelihood and Gaussian prior:

$$p(t \mid w) \propto \exp\Big(-\tfrac{\beta}{2}\|t-\Phi w\|^2\Big), \qquad p(w) \propto \exp\Big(-\tfrac{\alpha}{2}\|w\|^2\Big).$$

Posterior (unnormalized) by Bayes' rule:

$$p(w \mid t) \propto p(t \mid w)\, p(w) \propto \exp\Big(-\tfrac{\beta}{2}\|t-\Phi w\|^2 - \tfrac{\alpha}{2}\|w\|^2\Big).$$

Expand the exponents (quadratic form in w).

$$\begin{aligned} &\tfrac{\beta}{2}\|t-\Phi w\|^2 + \tfrac{\alpha}{2}\|w\|^2 \\ &= \tfrac{\beta}{2}\big(t^\top t - 2t^\top \Phi w + w^\top \Phi^\top \Phi w\big) + \tfrac{\alpha}{2} w^\top w \\ &= \tfrac{1}{2}\, w^\top(\beta\Phi^\top\Phi + \alpha I)\, w \;-\; \beta t^\top \Phi w + \tfrac{\beta}{2} t^\top t. \end{aligned}$$

Group terms in w and complete the square. Write the quadratic form as

$$\tfrac{1}{2}\, w^\top A\, w - b^\top w + \text{const}, \quad \text{where } A = \beta\Phi^\top\Phi + \alpha I, \quad b = \beta\Phi^\top t.$$

Complete the square:

$$\tfrac{1}{2} w^\top A w - b^\top w = \tfrac{1}{2}(w - A^{-1}b)^\top A (w - A^{-1}b) - \tfrac{1}{2} b^\top A^{-1} b.$$

Thus the unnormalized posterior becomes

$$p(w \mid t) \propto \exp\Big(-\tfrac{1}{2}(w-A^{-1}b)^\top A (w-A^{-1}b)\Big) \cdot \exp\Big(\tfrac{1}{2} b^\top A^{-1} b - \tfrac{\beta}{2} t^\top t\Big).$$

The second exponential is independent of w and becomes part of the normalizing constant.

Identify posterior covariance and mean. Hence the posterior is Gaussian with precision A and covariance $S_N = A^{-1}$:

$$S_N = (\beta\Phi^\top\Phi + \alpha I)^{-1},$$

and posterior mean

$$m_N = A^{-1}b = (\beta\Phi^\top\Phi + \alpha I)^{-1}(\beta\Phi^\top t).$$

Simplify using $\lambda = \alpha/\beta$. Dividing numerator and denominator by β gives the more familiar form:

$$S_N = \beta^{-1}(\Phi^\top\Phi + \lambda I)^{-1}, \qquad m_N = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top t,$$

where $\lambda = \alpha/\beta$. Note that m_N equals the ridge/MAP estimator and S_N quantifies posterior uncertainty.

Log-Likelihood and Log-Prior in Bayesian Linear Regression

Model Setup

We observe data $(\mathbf{X}, \mathbf{y})$ where $\mathbf{X} \in \mathbb{R}^{N\times D}$, $\mathbf{y} \in \mathbb{R}^N$ and weights $\mathbf{w} \in \mathbb{R}^D$. The linear-Gaussian model assumes

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}\big(0,\ \sigma^2 \mathbf{I}_N\big).$$

Log-Likelihood $\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$

Because the noise is i.i.d. Gaussian,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \mathcal{N}\big(\mathbf{y} \mid \mathbf{X}\mathbf{w},\ \sigma^2 \mathbf{I}_N\big).$$

Using the multivariate Gaussian density,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\right).$$

Thus the log-likelihood is

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}).$$

Equivalently,

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

Log-Prior $\log p(\mathbf{w})$

Assume a zero-mean Gaussian prior

$$p(\mathbf{w}) = \mathcal{N}\big(\mathbf{w} \mid 0,\ \alpha^{-1}\mathbf{I}_D\big).$$

The density is

$$p(\mathbf{w}) = \left(\frac{\alpha}{2\pi}\right)^{D/2} \exp\Big(-\frac{\alpha}{2}\,\mathbf{w}^\top\mathbf{w}\Big).$$

Therefore the log-prior is

$$\log p(\mathbf{w}) = \frac{D}{2}\log(\alpha) - \frac{D}{2}\log(2\pi) - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}.$$

MAP Estimation (Posterior Mode)

The posterior satisfies

$$p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})\, p(\mathbf{w}).$$

Hence,

$$\log p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}) + \text{const}.$$

Ignoring constants w.r.t. $\mathbf{w}$,

$$\log p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const}.$$

Maximizing the posterior is equivalent to minimizing

$$(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) \;+\; (\alpha\sigma^2)\,\mathbf{w}^\top\mathbf{w}.$$

Letting $\lambda = \alpha\sigma^2$, the MAP solution is the ridge-regression estimator

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}} \Big[\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2\Big].$$

LASSO (L1) as a MAP Estimate

We show that a Laplace prior on the weights leads to L1 regularization (LASSO).

1. The Laplace Prior Distribution

Assume a Laplace prior on the weights $\mathbf{w}$, which encourages sparsity (many weights at zero). We assume each weight w_j is drawn independently:

$$p(w_j) = \text{Laplace}(w_j \mid 0, b) = \frac{1}{2b} \exp\left(-\frac{|w_j|}{b}\right)$$

The full prior for the D-dimensional vector $\mathbf{w}$ is the product:

$$p(\mathbf{w}) = \prod_{j=1}^{D} p(w_j) = \left(\frac{1}{2b}\right)^D \exp\left(-\frac{1}{b}\sum_{j=1}^{D}|w_j|\right)$$

This can be written using the L1-norm, $\|\mathbf{w}\|_1 = \sum_{j=1}^{D}|w_j|$:

$$p(\mathbf{w}) = \left(\frac{1}{2b}\right)^D \exp\left(-\frac{1}{b}\|\mathbf{w}\|_1\right)$$

2. The Log-Prior

Taking the natural logarithm to get the log-prior:

$$\begin{aligned}\log p(\mathbf{w}) &= \log\left[\left(\frac{1}{2b}\right)^D \exp\left(-\frac{1}{b}\|\mathbf{w}\|_1\right)\right] \\ &= D\log\left(\frac{1}{2b}\right) - \frac{1}{b}\|\mathbf{w}\|_1 \\ &= \text{const} - \frac{1}{b}\|\mathbf{w}\|_1\end{aligned}$$

3. MAP Estimation

The MAP estimate maximizes the log-posterior, which is the sum of the log-likelihood and the log-prior:

$$\log p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}) + \text{const}$$

Substituting the Gaussian log-likelihood (from Section 7) [cite: 108] and the Laplace log-prior, ignoring all terms that are constant w.r.t. $\mathbf{w}$:

$$\log p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) \propto -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) - \frac{1}{b}\|\mathbf{w}\|_1$$

Maximizing this is equivalent to minimizing its negative. Using your defined '$\arg\min$' command:

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}}\left[\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \frac{1}{b}\|\mathbf{w}\|_1\right]$$

Multiplying by the constant $2\sigma^2$ and defining $\lambda = \frac{2\sigma^2}{b}$ gives the familiar LASSO objective function:

$$\boxed{\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}}\left[\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1\right]}$$

Generalized Least Squares (GLS)

1. The OLS Assumption vs. The GLS Model

The OLS estimator $\hat{w}_{\text{OLS}}$ is the BLUE (Best Linear Unbiased Estimator) under the Gauss-Markov assumptions, which critically require that the error covariance matrix is *spherical*:

$$\text{Cov}(\varepsilon) = \sigma^2 I_N$$

This single assumption implies two conditions:

- **Homoscedasticity:** All errors have the same variance σ^2.
- **No Autocorrelation:** All errors are uncorrelated.

In the **Generalized Least Squares (GLS)** model, we relax this assumption. We assume the errors are still zero-mean but have a general, known, $N \times N$ positive-definite covariance matrix $\mathbf{\Sigma}$:

$$\mathbb{E}[\varepsilon] = 0, \qquad \text{Cov}(\varepsilon) = \mathbf{\Sigma}$$

When $\mathbf{\Sigma} \neq \sigma^2 I_N$, the OLS estimator $\hat{w}_{\text{OLS}}$ is still *unbiased*, but it is no longer BLUE (i.e., it is not the minimum-variance estimator).

2. Derivation via Data Whitening

The core idea of GLS is to transform the generalized model back into a "standard" model that satisfies the OLS assumptions. This is done via *whitening*.

Since $\mathbf{\Sigma}$ is positive-definite, we can find a non-singular $N \times N$ matrix $\mathbf{C}$ such that $\mathbf{\Sigma} = \mathbf{C}\mathbf{C}^\top$ (e.g., via Cholesky decomposition). The inverse $\mathbf{C}^{-1}$ is our "whitening" matrix.

Start with the original model:

$$t = \Phi w + \varepsilon$$

Pre-multiply by $\mathbf{C}^{-1}$:

$$(\mathbf{C}^{-1}t) = (\mathbf{C}^{-1}\Phi)w + (\mathbf{C}^{-1}\varepsilon)$$

Let's define our transformed variables:

$$\tilde{t} = \mathbf{C}^{-1}t, \qquad \tilde{\Phi} = \mathbf{C}^{-1}\Phi, \qquad \tilde{\varepsilon} = \mathbf{C}^{-1}\varepsilon$$

Our new, transformed model is:

$$\tilde{t} = \tilde{\Phi}w + \tilde{\varepsilon}$$

Now, let's find the covariance of the *new* error term $\tilde{\varepsilon}$:

$$\begin{aligned}\text{Cov}(\tilde{\varepsilon}) &= \text{Cov}(\mathbf{C}^{-1}\varepsilon) \\ &= \mathbf{C}^{-1}\text{Cov}(\varepsilon)(\mathbf{C}^{-1})^\top \\ &= \mathbf{C}^{-1}\mathbf{\Sigma}(\mathbf{C}^\top)^{-1} \\ &= \mathbf{C}^{-1}(\mathbf{C}\mathbf{C}^\top)(\mathbf{C}^\top)^{-1} \\ &= (\mathbf{C}^{-1}\mathbf{C})(\mathbf{C}^\top(\mathbf{C}^\top)^{-1}) = I_N \cdot I_N = I_N\end{aligned}$$

The transformed model $\tilde{t} = \tilde{\Phi}w + \tilde{\varepsilon}$ has spherical errors ($\sigma^2 = 1$). It satisfies the Gauss-Markov assumptions!

3. The GLS (Aitken) Estimator

We can find the BLUE for w by simply applying the OLS formula to our *transformed* data:

$$\hat{w}_{\text{GLS}} = (\tilde{\Phi}^\top\tilde{\Phi})^{-1}\tilde{\Phi}^\top\tilde{t}$$

Now, substitute the original variables back in.

- $\tilde{\Phi}^\top\tilde{\Phi} = (\mathbf{C}^{-1}\Phi)^\top(\mathbf{C}^{-1}\Phi) = \Phi^\top(\mathbf{C}^{-1})^\top\mathbf{C}^{-1}\Phi = \Phi^\top(\mathbf{C}\mathbf{C}^\top)^{-1}\Phi = \Phi^\top\mathbf{\Sigma}^{-1}\Phi$
- $\tilde{\Phi}^\top\tilde{t} = (\mathbf{C}^{-1}\Phi)^\top(\mathbf{C}^{-1}t) = \Phi^\top(\mathbf{C}^{-1})^\top\mathbf{C}^{-1}t = \Phi^\top\mathbf{\Sigma}^{-1}t$

Substituting these gives the **GLS estimator**:

$$\boxed{\hat{w}_{\text{GLS}} = (\Phi^\top\mathbf{\Sigma}^{-1}\Phi)^{-1}\Phi^\top\mathbf{\Sigma}^{-1}t}$$

This is also called the **Aitken estimator**.

4. Properties of the GLS Estimator

Theorem 1.1.1 (Aitken Theorem)**.** *Under the generalized model $t = \Phi w + \varepsilon$ with $\text{Cov}(\varepsilon) = \mathbf{\Sigma}$:*

(i) **Unbiasedness:** *The GLS estimator is unbiased.*

$$\mathbb{E}[\hat{w}_{\text{GLS}}] = w$$

(ii) **Covariance:** *The covariance matrix of $\hat{w}_{\text{GLS}}$ is:*

$$\boxed{\text{Cov}(\hat{w}_{\text{GLS}}) = (\Phi^\top\mathbf{\Sigma}^{-1}\Phi)^{-1}}$$

(iii) **Efficiency:** *$\hat{w}_{\text{GLS}}$ is the* **BLUE** *(Best Linear Unbiased Estimator). Any other linear unbiased estimator $\tilde{w}$ will have a larger covariance.*

5. OLS as a Special Case of GLS

If the OLS assumptions were correct, $\mathbf{\Sigma} = \sigma^2 I_N$. Let's plug this into the GLS formula:

$$\begin{aligned}
\hat{w}_{\text{GLS}} &= (\Phi^\top(\sigma^2 I_N)^{-1}\Phi)^{-1}\Phi^\top(\sigma^2 I_N)^{-1}t \\
&= (\Phi^\top(\tfrac{1}{\sigma^2})\Phi)^{-1}\Phi^\top(\tfrac{1}{\sigma^2})t \\
&= \big(\tfrac{1}{\sigma^2}(\Phi^\top\Phi)\big)^{-1}\big(\tfrac{1}{\sigma^2}\Phi^\top t\big) \\
&= \big(\sigma^2(\Phi^\top\Phi)^{-1}\big)\big(\tfrac{1}{\sigma^2}\Phi^\top t\big) \\
&= (\Phi^\top\Phi)^{-1}\Phi^\top t = \hat{w}_{\text{OLS}}
\end{aligned}$$

This confirms that OLS is just a special case of GLS where the error covariance is spherical.

Note 1.1.2. **Feasible GLS (FGLS):** In practice, the exact covariance $\mathbf{\Sigma}$ is almost never known. **FGLS** is the practical approach where $\mathbf{\Sigma}$ is *estimated* from the data (often using the residuals from an initial OLS fit). The $\hat{\mathbf{\Sigma}}$ is then plugged into the GLS formula.

Derivation of GLS as an MLE

The GLS estimator is the MLE for a linear model where the noise is drawn from a single multivariate Gaussian distribution, allowing for both heteroscedasticity and autocorrelation.

1. Probabilistic Model & Error Function Assume the linear model in vector form:

$$\mathbf{t} = \mathbf{\Phi w} + \boldsymbol{\epsilon}$$

where the entire $N \times 1$ noise vector $\boldsymbol{\epsilon}$ is drawn from a zero-mean multivariate Gaussian with a general $N \times N$ positive-definite covariance matrix $\mathbf{\Sigma}$:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

This implies the likelihood for the entire target vector $\mathbf{t}$ is:

$$p(\mathbf{t} \mid \mathbf{\Phi}, \mathbf{w}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{t} \mid \mathbf{\Phi w}, \mathbf{\Sigma})$$

The probability density function (PDF) is:

$$p(\mathbf{t} \mid \dots) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t} - \mathbf{\Phi w})^\top\mathbf{\Sigma}^{-1}(\mathbf{t} - \mathbf{\Phi w})\right)$$

The log-likelihood $\mathcal{L}(\mathbf{w})$ is:

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{t} \mid \dots) = C - \frac{1}{2}(\mathbf{t} - \mathbf{\Phi w})^\top\mathbf{\Sigma}^{-1}(\mathbf{t} - \mathbf{\Phi w})$$

where $C = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}|$ is a constant with respect to $\mathbf{w}$.
To find the MLE, we maximize $\mathcal{L}(\mathbf{w})$, which is equivalent to minimizing the negative of the $\mathbf{w}$-dependent part. This gives the GLS error function $E(\mathbf{w})$:

$$E(\mathbf{w}) = (\mathbf{t} - \mathbf{\Phi w})^\top\mathbf{\Sigma}^{-1}(\mathbf{t} - \mathbf{\Phi w})$$

This quadratic form is the (generalized) squared Mahalanobis distance.

2. Derivation of the Closed-Form Solution The error function $E(\mathbf{w})$ is already in its matrix form. To find the minimum, we expand the expression. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ for simplicity.

$$E(\mathbf{w}) = (\mathbf{t}^\top - \mathbf{w}^\top\mathbf{\Phi}^\top)\mathbf{\Omega}(\mathbf{t} - \mathbf{\Phi w})$$

$$E(\mathbf{w}) = \mathbf{t}^\top\mathbf{\Omega t} - \mathbf{t}^\top\mathbf{\Omega\Phi w} - \mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Omega t} + \mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Omega\Phi w}$$

Since $\mathbf{\Sigma}$ is symmetric, its inverse $\mathbf{\Omega}$ is also symmetric ($\mathbf{\Omega}^\top = \mathbf{\Omega}$). The middle terms are transposes of each other:

$$E(\mathbf{w}) = \mathbf{t}^\top\mathbf{\Omega t} - 2\mathbf{w}^\top\mathbf{\Phi}^\top\mathbf{\Omega t} + \mathbf{w}^\top(\mathbf{\Phi}^\top\mathbf{\Omega\Phi})\mathbf{w}$$

Now, we take the gradient with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}}E(\mathbf{w}) = -2\mathbf{\Phi}^\top\mathbf{\Omega t} + 2(\mathbf{\Phi}^\top\mathbf{\Omega\Phi})\mathbf{w}$$

Set the gradient to zero to find the minimum:

$$\mathbf{0} = -2\mathbf{\Phi}^\top\mathbf{\Omega t} + 2(\mathbf{\Phi}^\top\mathbf{\Omega\Phi})\mathbf{w}$$

$$(\mathbf{\Phi}^\top\mathbf{\Omega\Phi})\mathbf{w} = \mathbf{\Phi}^\top\mathbf{\Omega t}$$

Substituting back $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, we get the **GLS Normal Equations**:

$$(\mathbf{\Phi}^\top\mathbf{\Sigma}^{-1}\mathbf{\Phi})\mathbf{w} = \mathbf{\Phi}^\top\mathbf{\Sigma}^{-1}\mathbf{t}$$

Assuming $(\mathbf{\Phi}^\top\mathbf{\Sigma}^{-1}\mathbf{\Phi})$ is invertible, we solve for $\mathbf{w}$ to get the GLS solution:

$$\boxed{\hat{\mathbf{w}}_{\text{GLS}} = (\mathbf{\Phi}^\top\mathbf{\Sigma}^{-1}\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}^{-1}\mathbf{t}}$$

Weighted Least Squares (WLS)

WLS is a special case of GLS used when errors are heteroscedastic but uncorrelated.

1. The WLS Model and Objective

We assume the general linear model $t = \Phi w + \varepsilon$, where $\mathbb{E}[\varepsilon] = 0$ but the errors are heteroscedastic:

$$\mathrm{Cov}(\varepsilon) = \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_N^2 \end{pmatrix}$$

The WLS objective is to minimize the **Weighted Sum of Squared Residuals (WSSR)**, where each squared residual is weighted by the inverse of its variance, $w_i = 1/\sigma_i^2$.

$$E(w) = \sum_{i=1}^{N} w_i (t_i - \phi_i^\top w)^2$$

In matrix form, we define the diagonal weight matrix $\mathbf{W} = \mathbf{\Sigma}^{-1}$:

$$\mathbf{W} = \mathrm{diag}(w_1, \dots, w_N) = \mathrm{diag}(1/\sigma_1^2, \dots, 1/\sigma_N^2)$$

The objective function becomes:

$$E(w) = (t - \Phi w)^\top \mathbf{W} (t - \Phi w)$$

2. Derivation of the WLS Estimator

We find the estimator $\hat{w}_{\mathrm{WLS}}$ by minimizing $E(w)$. First, expand the objective:

$$E(w) = t^\top \mathbf{W} t - t^\top \mathbf{W} \Phi w - w^\top \Phi^\top \mathbf{W} t + w^\top \Phi^\top \mathbf{W} \Phi w$$

Since $w^\top \Phi^\top \mathbf{W} t$ is a scalar, it equals its transpose $t^\top \mathbf{W} \Phi w$.

$$E(w) = t^\top \mathbf{W} t - 2 t^\top \mathbf{W} \Phi w + w^\top \Phi^\top \mathbf{W} \Phi w$$

Now, take the gradient with respect to w and set to zero:

$$\nabla_w E(w) = -2\Phi^\top \mathbf{W} t + 2\Phi^\top \mathbf{W} \Phi w$$

$$\nabla_w E(w) = 0 \quad \Rightarrow \quad 2\Phi^\top \mathbf{W} \Phi w = 2\Phi^\top \mathbf{W} t$$

This gives the **WLS Normal Equations**:

$$(\Phi^\top \mathbf{W} \Phi) w = \Phi^\top \mathbf{W} t$$

Assuming $\Phi^\top \mathbf{W} \Phi$ is invertible, the WLS estimator is:

$$\boxed{\hat{w}_{\mathrm{WLS}} = (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} t}$$

This is identical to the GLS estimator where $\mathbf{\Sigma}^{-1} = \mathbf{W}$.

3. Derivation via Data Whitening

We can also derive WLS by transforming the data so that the new error terms are homoscedastic with variance 1, and then applying OLS. Let $\mathbf{W}^{1/2}$ be the diagonal matrix with entries $\sqrt{w_i} = 1/\sigma_i$.

$$\mathbf{W}^{1/2} = \mathrm{diag}(1/\sigma_1, \dots, 1/\sigma_N)$$

Pre-multiply the original model $t = \Phi w + \varepsilon$ by $\mathbf{W}^{1/2}$:

$$(\mathbf{W}^{1/2} t) = (\mathbf{W}^{1/2} \Phi) w + (\mathbf{W}^{1/2} \varepsilon)$$

Define the transformed variables:

$$\tilde{t} = \mathbf{W}^{1/2} t, \qquad \tilde{\Phi} = \mathbf{W}^{1/2} \Phi, \qquad \tilde{\varepsilon} = \mathbf{W}^{1/2} \varepsilon$$

The new model is $\tilde{t} = \tilde{\Phi} w + \tilde{\varepsilon}$. Let's check the covariance of the new error $\tilde{\varepsilon}$:

$$\begin{aligned} \mathrm{Cov}(\tilde{\varepsilon}) &= \mathrm{Cov}(\mathbf{W}^{1/2} \varepsilon) \\ &= \mathbf{W}^{1/2} \mathrm{Cov}(\varepsilon) (\mathbf{W}^{1/2})^\top \\ &= \mathbf{W}^{1/2} \mathbf{\Sigma} \mathbf{W}^{1/2} \quad \text{(since } \mathbf{W} \text{ is diagonal)} \\ &= \mathbf{W}^{1/2} \mathbf{W}^{-1} \mathbf{W}^{1/2} \\ &= (\mathbf{W}^{1/2} \mathbf{W}^{-1/2})(\mathbf{W}^{-1/2} \mathbf{W}^{1/2}) = I \cdot I = I \end{aligned}$$

The transformed model has spherical errors, so we apply the OLS formula to it:

$$\hat{w} = (\tilde{\Phi}^\top \tilde{\Phi})^{-1} \tilde{\Phi}^\top \tilde{t}$$

Substitute the original variables back:

- $\tilde{\Phi}^\top \tilde{\Phi} = (\mathbf{W}^{1/2} \Phi)^\top (\mathbf{W}^{1/2} \Phi) = \Phi^\top (\mathbf{W}^{1/2})^\top \mathbf{W}^{1/2} \Phi = \Phi^\top \mathbf{W} \Phi$
- $\tilde{\Phi}^\top \tilde{t} = (\mathbf{W}^{1/2} \Phi)^\top (\mathbf{W}^{1/2} t) = \Phi^\top (\mathbf{W}^{1/2})^\top \mathbf{W}^{1/2} t = \Phi^\top \mathbf{W} t$

This yields the identical WLS estimator:

$$\hat{w}_{\mathrm{WLS}} = (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} t$$

4. Properties of the WLS Estimator

We assume the weights $\mathbf{W} = \mathbf{\Sigma}^{-1}$ are known.

Theorem 1.1.3 (Unbiasedness of WLS). *The WLS estimator is unbiased.*

Proof. Substitute $t = \Phi w + \varepsilon$ into the estimator:

$$\begin{aligned} \hat{w}_{\mathrm{WLS}} &= (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} (\Phi w + \varepsilon) \\ &= (\Phi^\top \mathbf{W} \Phi)^{-1} (\Phi^\top \mathbf{W} \Phi) w + (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} \varepsilon \\ &= w + (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} \varepsilon \end{aligned}$$

Now take the expectation:

$$\begin{aligned} \mathbb{E}[\hat{w}_{\mathrm{WLS}}] &= \mathbb{E}[w] + \mathbb{E}[(\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} \varepsilon] \\ &= w + (\Phi^\top \mathbf{W} \Phi)^{-1} \Phi^\top \mathbf{W} \mathbb{E}[\varepsilon] \\ &= w + 0 = w \end{aligned}$$

□

Theorem 1.1.4 (Covariance of WLS). *The covariance matrix of the WLS estimator is* $\mathrm{Cov}(\hat{w}_{\mathrm{WLS}}) = (\Phi^\top \mathbf{W} \Phi)^{-1}$.

Proof. Using the result from the unbiasedness proof:

$$\hat{w}_{\text{WLS}} - w = (\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top \mathbf{W}\varepsilon$$

Let $A = (\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top \mathbf{W}$. The covariance is:

$$\begin{aligned}
\text{Cov}(\hat{w}_{\text{WLS}}) &= \mathbb{E}[(\hat{w}_{\text{WLS}} - w)(\hat{w}_{\text{WLS}} - w)^\top] \\
&= \mathbb{E}[(A\varepsilon)(A\varepsilon)^\top] = \mathbb{E}[A\varepsilon\varepsilon^\top A^\top] \\
&= A\,\mathbb{E}[\varepsilon\varepsilon^\top]\,A^\top = A\,\mathbf{\Sigma}\,A^\top \\
&= \left[(\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top \mathbf{W}\right] \cdot \mathbf{\Sigma} \cdot \left[(\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top \mathbf{W}\right]^\top
\end{aligned}$$

Since $\mathbf{\Sigma} = \mathbf{W}^{-1}$ and $\mathbf{W}^\top = \mathbf{W}$ (it's diagonal):

$$\begin{aligned}
\text{Cov}(\hat{w}_{\text{WLS}}) &= \left[(\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top \mathbf{W}\right]\mathbf{W}^{-1}\left[\mathbf{W}\Phi(\Phi^\top \mathbf{W}\Phi)^{-1}\right] \\
&= (\Phi^\top \mathbf{W}\Phi)^{-1}\Phi^\top(\mathbf{W}\mathbf{W}^{-1})\mathbf{W}\Phi(\Phi^\top \mathbf{W}\Phi)^{-1} \\
&= (\Phi^\top \mathbf{W}\Phi)^{-1}(\Phi^\top \mathbf{W}\Phi)(\Phi^\top \mathbf{W}\Phi)^{-1} \\
&= I \cdot (\Phi^\top \mathbf{W}\Phi)^{-1} \\
&= \boxed{(\Phi^\top \mathbf{W}\Phi)^{-1}}
\end{aligned}$$

□

By the Aitken Theorem, $\hat{w}_{\text{WLS}}$ is the BLUE for this model.

5. Feasible WLS (FWLS)

In practice, the true variances σ_i^2 (and thus $\mathbf{W}$) are unknown. **Feasible WLS** is a multi-step process to estimate them:

1. Run OLS on $t = \Phi w + \varepsilon$ to get residuals $e = t - \Phi\hat{w}_{\text{OLS}}$.
2. Model the variance. Assume σ_i^2 is a function of some variables z_i (often $z_i = \phi_i$). To ensure positivity, model $\log(e_i^2)$:
$$\log(e_i^2) = z_i^\top \gamma + \nu_i$$
3. Use this model to predict the log-variances, then get predicted variances:
$$\hat{\sigma}_i^2 = \exp(z_i^\top \hat{\gamma})$$
4. Construct the estimated weight matrix: $\hat{\mathbf{W}} = \text{diag}(1/\hat{\sigma}_1^2, \ldots, 1/\hat{\sigma}_N^2)$.
5. Run WLS using $\hat{\mathbf{W}}$:
$$\hat{w}_{\text{FWLS}} = (\Phi^\top \hat{\mathbf{W}}\Phi)^{-1}\Phi^\top \hat{\mathbf{W}}t$$

This estimator is consistent and asymptotically more efficient than OLS.

Derivation of WLS as an MLE

We can derive the WLS estimator by finding the Maximum Likelihood Estimate (MLE) for a linear model with independent, heteroscedastic Gaussian noise.

1. Probabilistic Model & Error Function Assume the linear model $t_i = \phi_i^\top \mathbf{w} + \epsilon_i$, where the noise ϵ_i for each observation is drawn from a Gaussian with its own variance σ_i^2:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

The likelihood for a single observation t_i is:

$$p(t_i \mid \phi_i, \mathbf{w}, \sigma_i^2) = \mathcal{N}(t_i \mid \phi_i^\top \mathbf{w}, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(t_i - \phi_i^\top \mathbf{w})^2}{2\sigma_i^2}\right)$$

The log-likelihood for the entire dataset (assuming independence) is the sum:

$$\mathcal{L}(\mathbf{w}) = \log \prod_{i=1}^{N} p(t_i) = \sum_{i=1}^{N} \log p(t_i)$$

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma_i^2}}\right) - \frac{1}{2\sigma_i^2}(t_i - \phi_i^\top \mathbf{w})^2\right]$$

To find the MLE for $\mathbf{w}$, we maximize $\mathcal{L}(\mathbf{w})$. The first term in the sum is constant w.r.t. $\mathbf{w}$, so maximizing the log-likelihood is equivalent to minimizing the negative of the second term:

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(t_i - \phi_i^\top \mathbf{w})^2$$

Dropping the constant $1/2$ and defining the **weights** as $w_i = 1/\sigma_i^2$, we get the WLS error function $E(\mathbf{w})$:

$$E(\mathbf{w}) = \sum_{i=1}^{N} w_i(t_i - \phi_i^\top \mathbf{w})^2$$

2. Error Function in Matrix Form Let $\mathbf{t}$ be the $N \times 1$ target vector, $\mathbf{\Phi}$ the $N \times D$ design matrix, and $\mathbf{W}$ the $N \times N$ diagonal matrix of weights:

$$\mathbf{W} = \text{diag}(w_1, w_2, \ldots, w_N)$$

The error function $E(\mathbf{w})$ in matrix form is the quadratic form:

$$E(\mathbf{w}) = (\mathbf{t} - \mathbf{\Phi w})^\top \mathbf{W}(\mathbf{t} - \mathbf{\Phi w})$$

3. Derivation of the Closed-Form Solution To find the $\mathbf{w}$ that minimizes $E(\mathbf{w})$, we expand the expression and compute the gradient.

$$E(\mathbf{w}) = (\mathbf{t}^\top - \mathbf{w}^\top \mathbf{\Phi}^\top)\mathbf{W}(\mathbf{t} - \mathbf{\Phi w})$$

$$E(\mathbf{w}) = \mathbf{t}^\top \mathbf{W}\mathbf{t} - \mathbf{t}^\top \mathbf{W}\mathbf{\Phi w} - \mathbf{w}^\top \mathbf{\Phi}^\top \mathbf{W}\mathbf{t} + \mathbf{w}^\top \mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi w}$$

Since $\mathbf{W}$ is symmetric ($\mathbf{W}^\top = \mathbf{W}$), the two middle terms are transposes of each other (and are scalars), so we can combine them:

$$E(\mathbf{w}) = \mathbf{t}^\top \mathbf{W}\mathbf{t} - 2\mathbf{w}^\top \mathbf{\Phi}^\top \mathbf{W}\mathbf{t} + \mathbf{w}^\top (\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})\mathbf{w}$$

Now, we take the gradient with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = -2\mathbf{\Phi}^\top \mathbf{W}\mathbf{t} + 2(\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})\mathbf{w}$$

Set the gradient to zero to find the minimum:

$$\mathbf{0} = -2\mathbf{\Phi}^\top \mathbf{W}\mathbf{t} + 2(\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})\mathbf{w}$$

$$(\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})\mathbf{w} = \mathbf{\Phi}^\top \mathbf{W}\mathbf{t}$$

Assuming the matrix $(\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})$ is invertible, we solve for $\mathbf{w}$ to get the WLS solution:

$$\boxed{\hat{\mathbf{w}}_{\text{WLS}} = (\mathbf{\Phi}^\top \mathbf{W}\mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{W}\mathbf{t}}$$

Effects of Data Transformations on OLS Solution

We analyze the effect of common data operations on the OLS closed-form solution $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Scaling Individual Features

- **Operation:** Scale a single feature column j (assuming $j \geq 1$, not the bias) by a constant c.
- **Proof:** Let $\mathbf{C}$ be a diagonal matrix with $\mathbf{C}_{jj} = c$ and all other diagonal entries as 1. The new design matrix is $\mathbf{X}' = \mathbf{X}\mathbf{C}$.

$$\begin{aligned}\hat{\mathbf{w}}' &= ((\mathbf{X}\mathbf{C})^T(\mathbf{X}\mathbf{C}))^{-1}(\mathbf{X}\mathbf{C})^T\mathbf{y} \\ &= (\mathbf{C}^T\mathbf{X}^T\mathbf{X}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}^T\mathbf{y} \\ &= \mathbf{C}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{C}^T)^{-1}\mathbf{C}^T\mathbf{X}^T\mathbf{y} \\ &= \mathbf{C}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{C}^{-1}\hat{\mathbf{w}}\end{aligned}$$

- **Effect:** The new weight vector is $\hat{\mathbf{w}}' = \mathbf{C}^{-1}\hat{\mathbf{w}}$. Since $\mathbf{C}^{-1}$ is a diagonal matrix with $(\mathbf{C}^{-1})_{jj} = 1/c$, the corresponding weight $\hat{w}_j$ is scaled by $1/c$ ($\hat{w}'_j = \hat{w}_j/c$). All other weights, including the bias term $\hat{w}_0$, are unchanged.

Scaling All Features (not bias)

- **Operation:** Scale all feature columns $\mathbf{x}_j$ (for $j \geq 1$) by a constant c.
- **Proof:** This is the same as above, but $\mathbf{C} = \text{diag}(1, c, c, \ldots, c)$. The inverse is $\mathbf{C}^{-1} = \text{diag}(1, 1/c, 1/c, \ldots, 1/c)$. The proof $\hat{\mathbf{w}}' = \mathbf{C}^{-1}\hat{\mathbf{w}}$ is identical.
- **Effect:** The bias (intercept) term $\hat{w}_0$ is unchanged. All other feature weights $\hat{w}_j$ (for $j \geq 1$) are scaled by $1/c$.

Scaling Labels

- **Operation:** Scale the target vector $\mathbf{y}$ by a constant c. $\mathbf{y}' = c\mathbf{y}$.
- **Proof:**

$$\begin{aligned}\hat{\mathbf{w}}' &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}' \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(c\mathbf{y}) \\ &= c \cdot [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] = c \cdot \hat{\mathbf{w}}\end{aligned}$$

- **Effect:** All weights, including the bias term, are scaled by c.

Duplicating Rows

- **Operation:** Stack the entire dataset $(\mathbf{X}, \mathbf{y})$ on top of itself.
- **Proof:** The new matrices are $\mathbf{X}' = \begin{pmatrix}\mathbf{X} \\ \mathbf{X}\end{pmatrix}$ and $\mathbf{y}' = \begin{pmatrix}\mathbf{y} \\ \mathbf{y}\end{pmatrix}$.

$$\begin{aligned}(\mathbf{X}')^T\mathbf{X}' &= \begin{pmatrix}\mathbf{X}^T & \mathbf{X}^T\end{pmatrix}\begin{pmatrix}\mathbf{X} \\ \mathbf{X}\end{pmatrix} = 2(\mathbf{X}^T\mathbf{X}) \\ (\mathbf{X}')^T\mathbf{y}' &= \begin{pmatrix}\mathbf{X}^T & \mathbf{X}^T\end{pmatrix}\begin{pmatrix}\mathbf{y} \\ \mathbf{y}\end{pmatrix} = 2(\mathbf{X}^T\mathbf{y}) \\ \hat{\mathbf{w}}' &= (2(\mathbf{X}^T\mathbf{X}))^{-1}(2(\mathbf{X}^T\mathbf{y})) \\ &= \frac{1}{2}(\mathbf{X}^T\mathbf{X})^{-1}(2\mathbf{X}^T\mathbf{y}) = \hat{\mathbf{w}}\end{aligned}$$

- **Effect:** The solution $\hat{\mathbf{w}}$ is unchanged.

Removing Bias Term

- **Operation:** The original matrix $\mathbf{X} = [\mathbf{1}, \mathbf{X}_f]$ (where $\mathbf{X}_f$ are the features) becomes $\mathbf{X}' = \mathbf{X}_f$.
- **Proof:** The new solution $\hat{\mathbf{w}}' = (\mathbf{X}_f^T\mathbf{X}_f)^{-1}\mathbf{X}_f^T\mathbf{y}$ is not trivially related to the original $\hat{\mathbf{w}}$.
- **Effect:** The solution changes completely. The new model is forced to pass through the origin, which alters all coefficients.

Adding Dummy/Constant Features

- **Operation:** Add a new feature column that is constant, e.g., $\mathbf{x}_{\text{new}} = c \cdot \mathbf{1}$.
- **Proof:** The original matrix $\mathbf{X}$ already has a bias column (a column of 1s). The new column is a perfect linear combination of the bias column ($\mathbf{x}_{\text{new}} = c \cdot \mathbf{x}_0$). This is **perfect multicollinearity**.
- **Effect:** The columns of $\mathbf{X}'$ are linearly dependent, so the Gram matrix $(\mathbf{X}')^T\mathbf{X}'$ is singular (not invertible). A unique closed-form solution does not exist.

Duplicating Features

- **Operation:** Add a new feature column $\mathbf{x}_k$ that is identical to an existing column $\mathbf{x}_j$.
- **Proof:** The new column $\mathbf{x}_k$ is a perfect linear combination of $\mathbf{x}_j$ (i.e., $\mathbf{x}_k = 1 \cdot \mathbf{x}_j$). This is **perfect multicollinearity**.
- **Effect:** The Gram matrix $(\mathbf{X}')^T\mathbf{X}'$ is singular. A unique closed-form solution does not exist.

Adding a Single Data Row

- **Operation:** Add a new row $[\mathbf{x}_{\text{new}}^\top, y_{\text{new}}]$ to the dataset.
- **Proof:** The new matrices are $\mathbf{X}' = \begin{pmatrix}\mathbf{X} \\ \mathbf{x}_{\text{new}}^\top\end{pmatrix}$ and $\mathbf{y}' = \begin{pmatrix}\mathbf{y} \\ y_{\text{new}}\end{pmatrix}$.

$$\begin{aligned}(\mathbf{X}')^T\mathbf{X}' &= (\mathbf{X}^T\mathbf{X} + \mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}}^\top) \\ (\mathbf{X}')^T\mathbf{y}' &= (\mathbf{X}^T\mathbf{y} + \mathbf{x}_{\text{new}}y_{\text{new}}) \\ \hat{\mathbf{w}}' &= (\mathbf{X}^T\mathbf{X} + \mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}}^\top)^{-1}(\mathbf{X}^T\mathbf{y} + \mathbf{x}_{\text{new}}y_{\text{new}})\end{aligned}$$

- **Effect:** The solution $\hat{\mathbf{w}}$ changes. The new solution can be found from the old one using the Sherman-Morrison formula for rank-1 updates, but it is not a simple scaling.