Bayesian Linear Regression

Efim Abrikosov

January 7, 2024

1 Framework

1.1 Notations

$$y = x \cdot \boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{1.1}$$

1. Individual observations $(x, y) \in \mathbb{R}^k \times \mathbb{R}$
2. Observed data $(X, Y) \in \mathbb{R}^{n \times k} \times \mathbb{R}^n$
3. Linear regression weights $\boldsymbol{\beta} \in \mathbb{R}^k$
4. Observation error variance σ^2

1.2 Model Assumptions

1. Observations X have full rank
2. Observation errors are independent, normally distributed with mean zero and variance σ^2
3. Relation 1.1 holds
4. Error variance σ^2 is either known, or its prior distribution is inverse gamma distribution with parameters $a_0,\ b_0$
5. $\boldsymbol{\beta}$ has a conditional prior distribution $\beta \mid \sigma^2 \sim \mathcal{N}(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}_0)$

2 Known Observation Variance with Linear Weights Prior

2.1 Summary of Results

Posterior distribution of $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta} \mid Y, X) = \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2 \boldsymbol{\Sigma}_1) \tag{2.1}$$
$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^T X \tag{2.2}$$
$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_1 \left(X^T X \widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0 \right) \tag{2.3}$$
$$\widehat{\boldsymbol{\beta}} = \left(X^T X \right)^{-1} X^T Y \tag{2.4}$$

Posterior prediction distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta}$:

$$f(\widehat{y} \mid Y, X, x) = \mathcal{N} \left(x\boldsymbol{\beta}_1, \sigma^2 x \left(X^T X \right)^{-1} x^T \right) \tag{2.5}$$

Posterior observation distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta} + e$:

$$f(y \mid Y, X, x) = \mathcal{N} \left(x\boldsymbol{\beta}_1, \sigma^2 \left(1 + x \left(X^T X \right)^{-1} x^T \right) \right) \tag{2.6}$$

[Meaning?][Confidence interval, for a fixed value $\underline{\boldsymbol{\beta}}$ and a linear constraint $\boldsymbol{c} \in \mathbb{R}^k$:

$$\frac{\boldsymbol{c}(\underline{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)}{\sigma \sqrt{(\boldsymbol{c} \boldsymbol{\Sigma}_1 \boldsymbol{c}^T)}} \sim \mathcal{N}(0, 1) \tag{2.7}$$

Joint f-test for a set of linear constraints $\boldsymbol{C} \in \mathbb{R}^{l \times k}$]

2.2 Uninformative Prior

With the prior $\Lambda_0 \equiv \boldsymbol{\Sigma}_0^{-1} = 0$, the posterior 2.1 reduces to:

$$f(\boldsymbol{\beta} \mid Y, X) \sim \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2 \boldsymbol{\Sigma}_1) \tag{2.8}$$
$$\boldsymbol{\Sigma}_1 = \left(X^T X \right)^{-1} \tag{2.9}$$
$$\boldsymbol{\beta}_1 = \left(X^T X \right)^{-1} X^T Y \tag{2.10}$$

which is the standard result obtained in classical OLS set up.

The standard prediction interval for $\widehat{y}(x)$ and the confidence interval for an observation $y(x)$ follow from normal distributions in 2.5 and 2.6 respectively.

3 Conjugate Priors For Observation Variance and Linear Weights

3.1 Setup

$$f(Y, X \mid \boldsymbol{\beta}, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\boldsymbol{\beta})^T(y - X\boldsymbol{\beta})\right) \tag{3.1}$$

$$f(\boldsymbol{\beta} \mid \sigma^2) = |2\pi\boldsymbol{\Sigma}_0|^{-1} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\right) \tag{3.2}$$

$$f(\sigma^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-a_0-1} \exp\left(-\frac{b_0}{\sigma^2}\right) \tag{3.3}$$

Alternatively $f(\sigma^2)$ can be written as scaled inverse chi-squared distribution with parameters $(\nu_0, \tau_0^2) = \left(2a_0, \frac{b_0}{a_0}\right)$

3.2 Summary of Results

Posterior distribution of $\boldsymbol{\beta}$:

$$f(\sigma^2 \mid Y, X) = \text{Inv-}\Gamma(a_1, b_1) \tag{3.4}$$

$$a_1 = a_0 + \frac{n}{2} \tag{3.5}$$

$$b_1 = b_0 + \frac{1}{2}\left(Y^T Y + \boldsymbol{\beta}_0\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0^T - \boldsymbol{\beta}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\beta}_1^T\right) \tag{3.6}$$

$$f(\boldsymbol{\beta} \mid Y, X, \sigma^2) = \mathcal{N}(\boldsymbol{\beta}_1, \sigma^2\boldsymbol{\Sigma}_1) \tag{3.7}$$

$$\boldsymbol{\Sigma}_1^{-1} = \boldsymbol{\Sigma}_0^{-1} + X^T X \tag{3.8}$$

$$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_1\left(X^T X\widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right) \tag{3.9}$$

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y \tag{3.10}$$

Posterior prediction distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta}$:

$$f(\widehat{y} \mid Y, X, x) \propto \left(1 + \frac{a_1\left(y - x\boldsymbol{\beta}_1\right)^2}{vb_1}\frac{1}{2a_1}\right)^{-\frac{2a_1+1}{2}} \tag{3.11}$$

$$v = \left(1 - x\left(\boldsymbol{\Sigma}_1 + x^T x\right)^{-1} x^T\right)^{-1} \tag{3.12}$$

This is Student's t-distribution on $y - x\boldsymbol{\beta}_1$ with scale $\frac{vb_1}{a_1}$ and $2a_1$ degrees of freedom.

Posterior observation distribution $\widehat{y} \equiv x \cdot \boldsymbol{\beta} + e$:

$$f(y \mid Y, X, x) = ?? \tag{3.13}$$