

STAT 238 - Bayesian Statistics

Lecture Nineteen

Spring 2026, UC Berkeley

Aditya Guntuboyina

09 March 2026

1 A High-Dimensional Linear Regression Model

We have a response variable y and a single covariate x . In our example, y denotes weekly earnings and x denotes years of experience. The covariate x takes the values $0, 1, \dots, m$ for some fixed integer m .

Our data is $(x_i, y_i), i = 1, \dots, n$. The model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \text{ReLU}(x_i - 1) + \dots + \beta_m \text{ReLU}(x_i - (m - 1)) + \epsilon_i \quad (1)$$

where $\text{ReLU}(u) := \max(u, 0)$, and $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Note we did not include $\text{ReLU}(x_i - m)$ because it always equals 0.

This linear regression equation can be written in the usual notation as:

$$y = X\beta + \epsilon \quad (2)$$

where $y = (y_1, \dots, y_n)^T$ and X is the $n \times (m + 1)$ matrix with columns 1, x , $\text{ReLU}(x - j)$ for $j = 1, \dots, m - 1$.

2 Recap: Linear Regression with default priors

As we saw before, the default prior on β_0, \dots, β_m in the linear regression model (2) is:

$$\beta_0, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} \text{uniform}(-C, C) \quad (3)$$

with $C \rightarrow \infty$. With this prior, the posterior of β given σ becomes (see e.g., Problem 7(a) in Homework Three)

$$\beta \mid \text{data}, \sigma \sim N((X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1}) \quad (4)$$

which means that the posterior mean of β is just the least squares estimator $(X^T X)^{-1} X^T y$ (note that this posterior mean of β does not depend on σ , so is both the conditional posterior mean given σ , as well as the unconditional posterior mean).

The fact that the posterior mean of β coincides with the least squares estimator is shared by some priors other than (3) as well. For example, this is also true for the following prior:

$$\beta_0, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} N(0, C) \quad (5)$$

as $C \rightarrow \infty$. This can be seen by the following general result (which follows, for example, from Problem 9 in Homework Two):

$$\beta \mid \sigma \sim N(0, Q) \implies \beta \mid \text{data}, \sigma \sim N \left(\left(\frac{X^T X}{\sigma^2} + Q^{-1} \right)^{-1} \frac{X^T y}{\sigma^2}, \left(\frac{X^T X}{\sigma^2} + Q^{-1} \right)^{-1} \right). \quad (6)$$

From here, it is clear that if Q is chosen so that $Q^{-1} \approx 0$, then we get the same posterior as (4). The prior (5) corresponds to $Q = CI$ so that $Q^{-1} = (1/C)I$ which goes to zero when $C \rightarrow \infty$. So the prior (5) leads to the same posterior (4).

3 Regularization

In the earnings-experience regression example, we have seen in the last class that the least squares estimator for (2) is not interpretable. We want the fitted curve

$$x \mapsto \hat{\beta}_0 + \hat{\beta}_1 x + \sum_{j=2}^m \hat{\beta}_j \text{ReLU}(x - (j-1))$$

to be smooth for it to be interpretable. This can be achieved in the Bayesian setting if we use the prior:

$$\beta_0, \beta_1 \stackrel{\text{i.i.d.}}{\sim} N(0, C) \quad \text{and} \quad \beta_2, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} N(0, \tau^2). \quad (7)$$

for some small τ . The above prior is equivalent to $\beta \sim N(0, Q)$ where Q is the $(m+1) \times (m+1)$ diagonal matrix with diagonal entries $C, C, \tau^2, \dots, \tau^2$. So using the formula (6), the posterior mean corresponding to this prior is:

$$\mathbb{E}(\beta \mid \text{data}, \sigma, \tau) = \left(\frac{X^T X}{\sigma^2} + Q^{-1} \right)^{-1} \frac{X^T y}{\sigma^2}.$$

Because Q is diagonal with diagonal entries $C, C, 1/\tau^2, \dots, 1/\tau^2$, it is easy to see that Q^{-1} is also diagonal with diagonal entries $1/C, 1/C, \tau^{-2}, \dots, \tau^{-2}$. Because C is large, we can approximate Q^{-1} by the matrix with diagonal entries $0, 0, \tau^{-2}, \dots, \tau^{-2}$. In other words:

$$Q^{-1} \approx \frac{1}{\tau^2} J \quad \text{where} \quad J = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

Note that J is an $(m+1) \times (m+1)$ diagonal matrix. Thus the posterior mean becomes:

$$\mathbb{E}(\beta \mid \text{data}, \sigma, \tau) = \left(\frac{X^T X}{\sigma^2} + \frac{J}{\tau^2} \right)^{-1} \frac{X^T y}{\sigma^2} = \left(X^T X + \frac{J}{\tau^2/\sigma^2} \right)^{-1} X^T y.$$

We use the notation $\tau^2 = \sigma^2 \gamma^2$ or equivalently $\gamma^2 = \tau^2/\sigma^2$. This gives

$$\mathbb{E}(\beta \mid \text{data}, \sigma, \tau) = \left(X^T X + \frac{J}{\gamma^2} \right)^{-1} X^T y. \quad (8)$$

It can be verified that this quantity $(X^T X + J/\gamma^2)^{-1} X^T y$ coincides with a ridge regularized least squares estimator. This is shown in the next section.

4 Connection to Ridge Regularization

For the model (1), the ridge regression estimator $\hat{\beta}_{\text{ridge}}(\lambda)$ for β is given by the minimizer of:

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i - \beta_2 \text{ReLU}(x_i - 1) - \dots - \beta_m \text{ReLU}(x_i - (m - 1)))^2 + \lambda (\beta_2^2 + \beta_3^2 + \dots + \beta_m^2). \quad (9)$$

The objective function above can also be written as

$$\|y - X\beta\|^2 + \lambda \sum_{j=2}^m \beta_j^2.$$

It turns out that $\hat{\beta}_{\text{ridge}}(\lambda)$ can be written in closed form using matrix notation. To see this, note first that the gradient of the above objective function with respect to β is given by

$$\nabla \left(\|y - X\beta\|^2 + \lambda \sum_{j=2}^m \beta_j^2 \right) = -2X^T y + 2X^T X\beta + 2\lambda \begin{pmatrix} 0 \\ 0 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix}$$

Recalling that J is the $(m + 1) \times (m + 1)$ diagonal matrix with diagonal entries $0, 0, 1, \dots, 1$, we can write

$$\nabla \left(\|y - X\beta\|^2 + \lambda \sum_{t=2}^{n-1} \beta_t^2 \right) = -2X^T y + 2X^T X\beta + 2\lambda \begin{pmatrix} 0 \\ 0 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_{n-1} \end{pmatrix} = -2X^T y + 2X^T X\beta + 2\lambda J\beta.$$

Setting this gradient equal to zero, we get

$$-2X^T y + 2X^T X\beta + 2\lambda J\beta = 0 \implies (X^T X + \lambda J)\beta = X^T y.$$

which gives

$$\hat{\beta}_{\text{ridge}}(\lambda) = (X^T X + \lambda J)^{-1} X^T y. \quad (10)$$

Note that $\lambda = 0$ corresponds to least squares, and $\lambda \rightarrow \infty$ leads to linear regression. When λ is set to be neither very close to 0 nor very large, we should get a smooth estimate that still gives a good fit to the data.

Comparing (10) to (8), it is clear that they are identical when $\lambda = 1/\gamma^2$. In other words, when we use the prior (7), then the posterior mean (given τ and σ) coincides with ridge regularized least squares with regularization parameter $\lambda = 1/\gamma^2 = \sigma^2/\tau^2$.

5 Bayesian inference for γ and σ

The big advantage of Bayesian inference is that it also allows inference on γ and σ . We shall use the prior:

$$\log \gamma, \log \sigma \stackrel{\text{i.i.d}}{\sim} \text{uniform}(-\infty, \infty).$$

Recall again that $\tau = \gamma\sigma$. The joint prior on all the parameters (β, γ, σ) is:

$$f_{\beta, \gamma, \sigma}(\beta, \gamma, \sigma) \propto \frac{1}{\gamma\sigma} \frac{1}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}\beta^T Q^{-1}\beta\right).$$

Here Q is the $(m+1) \times (m+1)$ diagonal matrix with diagonal entries $C, C, \gamma^2\sigma^2, \dots, \gamma^2\sigma^2$.

With this prior, we shall argue in the next lecture that the posterior is given by:

$$\beta \mid \text{data}, \sigma, \gamma \sim N\left((X^T X + \gamma^{-2}J)^{-1} X^T y, \sigma^2 (X^T X + \gamma^{-2}J)^{-1}\right) \quad (11)$$

and

$$\frac{1}{\sigma^2} \mid \text{data}, \gamma \sim \text{Gamma}\left(\frac{n}{2} - 1, \frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2}\right) \quad (12)$$

and

$$f_{\gamma \mid \text{data}}(\gamma) \propto \frac{\gamma^{-m} |X^T X + \gamma^{-2}J|^{-1/2}}{\left(y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y\right)^{(n/2)-1}}.$$

With this, inference can be carried out by first taking a grid of γ values and computing the above posterior (on the logarithmic scale) at the grid points. This posterior can be used to obtain posterior samples of γ . For each sample of γ , we can then sample σ using the distribution (12). Given samples from both γ and σ , we can then sample β using (11).