

STA 360/602L: MODULE 6.1

BAYESIAN LINEAR REGRESSION

DR. OLANREWaju MICHAEL AKANDE

MOTIVATING EXAMPLE

- Let's consider the problem of predicting swimming times for high school swimmers to swim 50 yards.
- We have data collected on four students, each with six times taken (every two weeks).
- Suppose the coach of the team wants to use the data to recommend one of the swimmers to compete in a swim meet in two weeks time.
- Since we want to predict swimming times given week, one option would be regression models.
- In a typical regression setup, we store the predictor variables in a matrix $\mathbf{X}_{n \times p}$, so n is the number of observations and p is the number of variables.
- You should all know how to write down and fit linear regression models of the most common forms, so let's only review the most important details.

NORMAL REGRESSION MODEL

- The model assumes the following distribution for a response variable Y_i given multiple covariates/predictors $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{i(p-1)})$.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{i(p-1)} + \epsilon_i; \quad \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

or in vector form for the parameters,

$$Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i; \quad \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1})$.

- We can also write the model as:

$$Y_i \stackrel{iid}{\sim} \mathcal{N}(\boldsymbol{\beta}^T \mathbf{x}_i, \sigma^2);$$

$$p(y_i | \mathbf{x}_i) = \mathcal{N}(\boldsymbol{\beta}^T \mathbf{x}_i, \sigma^2).$$

- That is, the model assumes $\mathbb{E}[Y | \mathbf{x}]$ is linear.

LIKELIHOOD

- Given that we have $Y_i \stackrel{iid}{\sim} \mathcal{N}(\beta^T \mathbf{x}_i, \sigma^2)$, the likelihood is

$$\begin{aligned} p(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n, \beta, \sigma^2) &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \beta, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \beta^T \mathbf{x}_i)^2 \right\} \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 \right\}. \end{aligned}$$

- From all our work with normal models, we already know it would be convenient to specify a (multivariate) normal prior on β and a gamma prior on $1/\sigma^2$, so let's start there.
- Two things to immediately notice:
 - since β is a vector, it might actually be better to rewrite this kernel in multivariate form altogether, and
 - when combining this likelihood with the prior kernel, we will need to find a way to detach β from \mathbf{x}_i .

MULTIVARIATE FORM

- Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

- Then, we can write the model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_{n \times n}).$$

- That is, in multivariate form, we have

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n \times n}).$$

FREQUENTIST ESTIMATION RECAP

- OLS estimate of β is given by

$$\hat{\beta}_{\text{ols}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- Predictions can then be written as

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\beta}_{\text{ols}} = \mathbf{X} \left[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \right] = \left[\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \right] \mathbf{y}.$$

- The variance of the OLS estimates of all p coefficients is

$$\text{Var} \left[\hat{\beta}_{\text{ols}} \right] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}.$$

- Finally,

$$s_e^2 = \frac{(\mathbf{y} - \mathbf{X} \hat{\beta}_{\text{ols}})^T (\mathbf{y} - \mathbf{X} \hat{\beta}_{\text{ols}})}{n - p}.$$

BAYESIAN SPECIFICATION

- The likelihood for the regression model becomes

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \right\} \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} [\mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta] \right\}. \end{aligned}$$

- We can start with the following semi-conjugate prior for β :

$$\pi(\beta) = \mathcal{N}_p(\beta_0, \Sigma_0).$$

- That is, the pdf is

$$\pi(\beta) = (2\pi)^{-\frac{p}{2}} |\Sigma_0|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right\}.$$

- Recall from our multivariate normal model that we can write this pdf as

$$\pi(\beta) \propto \exp \left\{ -\frac{1}{2} \beta^T \Sigma_0^{-1} \beta + \beta^T \Sigma_0^{-1} \mu_0 \right\}.$$

MULTIVARIATE NORMAL MODEL RECAP

- To avoid doing all work from scratch, we can leverage results from the multivariate normal model.
- In particular, recall that if $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$,

$$p(\mathbf{y}|\boldsymbol{\theta}, \Sigma) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T (\Sigma^{-1}) \boldsymbol{\theta} + \boldsymbol{\theta}^T (\Sigma^{-1} \bar{\mathbf{y}}) \right\}$$

and

$$\pi(\boldsymbol{\theta}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\theta} + \boldsymbol{\theta}^T \Lambda_0^{-1} \boldsymbol{\mu}_0 \right\}$$

- Then

$$\pi(\boldsymbol{\theta}|\Sigma, \mathbf{y}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\theta}^T [\Lambda_0^{-1} + \Sigma^{-1}] \boldsymbol{\theta} + \boldsymbol{\theta}^T [\Lambda_0^{-1} \boldsymbol{\mu}_0 + \Sigma^{-1} \bar{\mathbf{y}}] \right\} \equiv \mathcal{N}_p(\boldsymbol{\mu}_n, \Lambda_n)$$

where

$$\begin{aligned} \Lambda_n &= [\Lambda_0^{-1} + \Sigma^{-1}]^{-1} \\ \boldsymbol{\mu}_n &= \Lambda_n [\Lambda_0^{-1} \boldsymbol{\mu}_0 + \Sigma^{-1} \bar{\mathbf{y}}] . \end{aligned}$$

POSTERIOR COMPUTATION

- For inference on β , rewrite the likelihood as

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} [\mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta] \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} [\beta^T \mathbf{X}^T \mathbf{X} \beta - 2\beta^T \mathbf{X}^T \mathbf{y}] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \beta^T \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} \right) \beta + \beta^T \left(\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y} \right) \right\}. \end{aligned}$$

- Again, with the prior written as

$$\pi(\beta) \propto \exp \left\{ -\frac{1}{2} \beta^T \Sigma_0^{-1} \beta + \beta^T \Sigma_0^{-1} \mu_0 \right\},$$

both forms look like what we have on the previous page. It is then easy to read off the full conditional for β .

POSTERIOR COMPUTATION

- That is,

$$\begin{aligned}\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2) \cdot \pi(\boldsymbol{\beta}) \\ &\propto \exp \left\{ -\frac{1}{2} \boldsymbol{\beta}^T \left[\Sigma_0^{-1} + \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} \right] \boldsymbol{\beta} + \boldsymbol{\beta}^T \left[\Sigma_0^{-1} \boldsymbol{\beta}_0 + \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y} \right] \right\} \\ &\equiv \mathcal{N}_p(\boldsymbol{\mu}_n, \Sigma_n).\end{aligned}$$

- Comparing this to the prior

$$\pi(\boldsymbol{\beta}) \propto \exp \left\{ -\frac{1}{2} \boldsymbol{\beta}^T \Sigma_0^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}^T \Sigma_0^{-1} \boldsymbol{\mu}_0 \right\},$$

means

$$\begin{aligned}\Sigma_n &= \left[\Sigma_0^{-1} + \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X} \right]^{-1} \\ \boldsymbol{\mu}_n &= \Sigma_n \left[\Sigma_0^{-1} \boldsymbol{\beta}_0 + \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{y} \right].\end{aligned}$$

POSTERIOR COMPUTATION

- Next, we move to σ^2 . From previous work, we already know the inverse-gamma distribution will be semi-conjugate.

- First, recall that $\mathcal{IG}(y; a, b) \equiv \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}}$.

- So, if we set $\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$, we have

$$\pi(\sigma^2 | \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) \propto p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \sigma^2) \cdot \pi(\sigma^2)$$

$$\begin{aligned} &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ - \left(\frac{1}{\sigma^2} \right) \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} \right\} \\ &\quad \times (\sigma^2)^{-\left(\frac{\nu_0}{2} + 1\right)} e^{-\left(\frac{1}{\sigma^2}\right) \left[\frac{\nu_0 \sigma_0^2}{2} \right]} \end{aligned}$$

POSTERIOR COMPUTATION

- That is,

$$\begin{aligned}\pi(\sigma^2 | \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}) &\propto (\sigma^2)^{-\frac{n}{2}} \exp \left\{ - \left(\frac{1}{\sigma^2} \right) \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} \right\} \\ &\quad \times (\sigma^2)^{-\left(\frac{\nu_0}{2} + 1\right)} e^{-\left(\frac{1}{\sigma^2}\right) \left[\frac{\nu_0 \sigma_0^2}{2} \right]} \\ &\propto (\sigma^2)^{-\left(\frac{\nu_0 + n}{2} + 1\right)} e^{-\left(\frac{1}{\sigma^2}\right) \left[\frac{\nu_0 \sigma_0^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2} \right]} \\ &\equiv \mathcal{IG} \left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2} \right),\end{aligned}$$

where

$$\nu_n = \nu_0 + n; \quad \sigma_n^2 = \frac{1}{\nu_n} [\nu_0 \sigma_0^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] = \frac{1}{\nu_n} [\nu_0 \sigma_0^2 + \text{SSR}(\boldsymbol{\beta})].$$

- $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is the sum of squares of the residuals (SSR).

WHAT'S NEXT?

MOVE ON TO THE READINGS FOR THE NEXT MODULE!