STA 360/602L: Module 6.1

Bayesian linear regression

Dr. Olanrewaju Michael Akande

STA 602L

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 $\boldsymbol{X}_{n \times p}$, 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 $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{i(p-1)})$.

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

  or in vector form for the parameters,

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

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

- We can also write the model as:

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

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

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

Likelihood

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

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

- From all our work with normal models, we already know it would be convenient to specify a (multivariate) normal prior on $\boldsymbol{\beta}$ and a gamma prior on $1/\sigma^2$, so let's start there.

- Two things to immediately notice:
  - since $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ from $\boldsymbol{x}_i$.

STA 602L

Multivariate form

- Let

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

- Then, we can write the model as

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

- That is, in multivariate form, we have

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

STA 602L

Frequentist estimation recap

- OLS estimate of $\boldsymbol{\beta}$ is given by

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

- Predictions can then be written as

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

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

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

- Finally,

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

STA 602L

Bayesian specification

- The likelihood for the regression model becomes

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

- We can start with the following semi-conjugate prior for $\boldsymbol{\beta}$:

$$\pi(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_0, \Sigma_0).$$

- That is, the pdf is

$$\pi(\boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}}|\Sigma_0|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\}.$$

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

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

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 $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma)$,

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

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

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

where

$$\Lambda_n = \left[\Lambda_0^{-1} + \Sigma^{-1}\right]^{-1}$$
$$\boldsymbol{\mu}_n = \Lambda_n\left[\Lambda_0^{-1}\boldsymbol{\mu}_0 + \Sigma^{-1}\bar{\boldsymbol{y}}\right].$$

Posterior computation

- For inference on $\boldsymbol{\beta}$, rewrite the likelihood as

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

- Again, with the prior written as

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

both forms look like what we have on the previous page. It is then easy to read off the full conditional for $\boldsymbol{\beta}$.

STA 602L

Posterior computation

- That is,

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

- 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

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

Posterior computation

- Next, we move to σ^2. From previous work, we already know the inverse-gamma distribution with 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

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

STA 602L

Posterior computation

- That is,

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

where

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

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

STA 602L

What's next?

Move on to the readings for the next module!

STA 602L