An Introduction to Bayesian Linear Regression

APPM 5720: Bayesian Computation

Fall 2018

A Simple Linear Model

Suppose that we observe

- explanatory variables $x_1, x_2, \ldots, x_n$

and

- dependent variables $y_1, y_2, \ldots, y_n$

Assume they are related through the very simple linear model

$$y_i = \beta x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$, with $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ being realizations of iid $N(0, \sigma^2)$ random variables.

A Simple Linear Model

$$y_i = \beta x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

- The x_i can either be constants or realizations of random variables.
- In the latter case, assume that they have joint pdf $f(\vec{x}|\theta)$ where θ is a parameter (or vector of parameters) that is unrelated to β and σ^2.

The likelihood for this model is

$$\begin{aligned} f(\vec{y}, \vec{x}|\beta, \sigma^2, \theta) &= f(\vec{y}|\vec{x}, \beta, \sigma^2, \theta) \cdot f(\vec{x}|\beta, \sigma^2, \theta) \\ &= f(\vec{y}|\vec{x}, \beta, \sigma^2) \cdot f(\vec{x}|\theta) \end{aligned}$$

A Simple Linear Model

- Assume that the x_i are fixed. The likelihood for the model is then $f(\vec{y}|\vec{x}, \beta, \sigma^2)$.

- The goal is to estimate and make inferences about the parameters β and σ^2.

Frequentist Approach: Ordinary Least Squares (OLS)

- y_i is supposed to be β times x_i plus some residual noise.
- The noise, modeled by a normal distribution, is observed as $y_i - \beta x_i$.
- Take β to be the minimizer of the sum of squared errors

$$\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

A Simple Linear Model

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Now for the randomness. Consider

$$Y_i = \beta x_i + Z_i, \qquad i = 1, 2, \ldots, n$$

for $Z_i \overset{iid}{\sim} N(0, \sigma^2)$.
Then

- $Y_i \sim N(\beta x_i, \sigma^2)$
- $$\widehat{\beta} = \sum_{i=1}^{n} \left(\frac{x_i}{\sum x_j^2} \right) Y_i \sim N\left(\beta, \sigma^2 / \sum x_j^2\right)$$

A Simple Linear Model

If we predict each y_i to be $\widehat{y}_i := \widehat{\beta} x_i$, we can define the sum of squared errors to be

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta} x_i)^2$$

We can then estimate the noise variance σ^2 by the average sum of squared errors SSE/n or, better yet, we can adjust the denominator slightly to get the unbiased estimator

$$\widehat{\sigma^2} = \frac{SSE}{n-1}.$$

This quantity is known as the mean squared error or MSE and will also be denoted by s^2.

The Bayesian Approach:

$$Y_i = \beta x_i + Z_i, \qquad Z_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

$$\Rightarrow \; f(y_i | \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2\right]$$

$$\Rightarrow \; f(\vec{y} | \beta, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2\right]$$

It will be convenient to write this in terms of the OLS estimators

$$\widehat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}, \qquad s^2 = \frac{\sum (y_i - \widehat{\beta} x_i)^2}{n-1}$$

THE BAYESIAN APPROACH:

Then

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2 = \nu s^2 + (\beta - \widehat{\beta})^2 \sum_{i=1}^{n} x_i^2$$

where $\nu := n - 1$.

It will also be convenient to work with the precision parameter $\tau := 1/\sigma^2$.

Then

$$\begin{aligned} f(\vec{y}|\beta, \tau) &= (2\pi)^{-n/2} \\ &\cdot \left\{ \tau^{1/2} \cdot \exp\left[-\tfrac{\tau}{2} (\beta - \widehat{\beta})^2 \textstyle\sum_{i=1}^{n} x_i^2 \right] \right\} \\ &\cdot \left\{ \tau^{\nu/2} \cdot \exp\left[-\tfrac{\tau \nu s^2}{2} \right] \right\} \end{aligned}$$

THE BAYESIAN APPROACH:

- $\tau^{1/2} \cdot \exp\left[-\frac{\tau}{2}(\beta - \widehat{\beta})^2 \sum_{i=1}^{n} x_i^2\right]$

  looks normal as a function of β

- $\tau^{\nu/2} \cdot \exp\left[-\frac{\tau \nu s^2}{2}\right]$

  looks gamma as a function of τ

  (inverse gamma as a function of σ^2)

The natural conjugate prior for (β, σ^2) will be a "normal inverse gamma".

The Bayesian Approach:

So many symbols... will use “underbars” and “overbars” for prior and posterior hyperparameters and also add a little more structure.

- Priors

$$\beta|\tau \sim N(\underline{\beta}, \underline{c}/\tau), \qquad \tau \sim \Gamma(\underline{\nu}/2, \underline{\nu}\,\underline{s}^2/2)$$

- Will write

$$(\beta, \tau) \sim NG(\underline{\beta}, c, \underline{\nu}/2, \underline{\nu}\,\underline{s}^2/2).$$

The Bayesian Approach:

It is “routine” to show that the posterior is

$$(\beta, \tau)|\vec{y} \sim NG(\overline{\beta}, \overline{c}, \overline{\nu}/2, \overline{\nu}\,\overline{s^2}/2)$$

where

$$\overline{c} = \left[1/\underline{c} + \sum x_i^2\right]^{-1}, \qquad \overline{\beta} = \overline{c}(\underline{c}^{-1}\underline{\beta} + \widehat{\beta}\sum x_i^2)$$

$$\overline{\nu} = \underline{\nu} + n, \qquad \overline{\nu}\overline{s^2} = \underline{\nu s^2} + \nu s^2 + \frac{(\widehat{\beta} - \underline{\beta})^2}{\underline{c} + \sum x_i^2}$$

ESTIMATING β AND σ^2:

- The posterior Bayes estimator for β is $\mathsf{E}[\beta|\vec{y}]$.
- A measure of uncertainty of the estimator is given by the posterior variance $Var[\beta|\vec{y}]$.
- We need to write down the $NG(\overline{\beta}, \overline{c}, \overline{\nu}/2, \overline{\nu}\,\overline{s^2}/2)$ pdf for $(\beta, \tau)|\vec{y}$ and integrate out τ.
- The result is that $\beta|\vec{y}$ has a generalized t-distribution. (This is not exactly the same as a non-central t.)

The Multivariate t-Distribution:

We say that a k-dimensional random vector $\vec{X}$ has a multivariate t-distribution with

- mean $\vec{\mu}$
- variance-covariance matrix parameter V
- ν degrees of freedon

if $\vec{X}$ has pdf

$$f(\vec{x}|\vec{\mu}, V, \nu) = \frac{\nu^{\nu/2}\Gamma\left(\frac{\nu+k}{2}\right)}{\pi^{k/2}\Gamma\left(\frac{\nu}{2}\right)}|V|^{-1/2}\left[(\vec{x}-\vec{\mu})^t V^{-1}(\vec{x}-\vec{\mu})+\nu\right]^{-\frac{\nu+k}{2}}.$$

We will write

$$\vec{X} \sim t(\vec{\mu}, V, \nu).$$

The Multivariate t-Distribution:

- With $k = 1$, $\vec{\mu} = 0$, and $V = 1$, we get the usual t-distribution.
- Marginals:

$$\vec{X} = \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix} \qquad \Rightarrow \qquad \vec{X}_i \sim t(\vec{\mu}_i, V_i, \nu)$$

  where $\vec{\mu}_i$ and V_i are the mean and variance-covariance matrix of $\vec{X}_i$.
- Conditionals such as $\vec{X}_1|\vec{X}_2$ are also multivariate t.
- $$\begin{aligned} \mathsf{E}[\vec{X}] &= \vec{\mu}, \text{ if } \nu > 1 \\ Var[\vec{X}] &= \tfrac{\nu}{\nu-2} V \text{ if } \nu > 2 \end{aligned}$$

Back to the Regression Problem:

- Can show that $\beta|\vec{y} \sim t(\overline{\beta}, \overline{cs^2}, \overline{\nu})$
  So, the PBE is
  $$\mathsf{E}[\beta|\vec{y}] = \overline{\beta}$$
  and the posterior variance is
  $$Var[\beta|\vec{y}] = \frac{\overline{\nu}}{\overline{\nu} - 1}\overline{cs^2}.$$
- Also can show that $\tau|\vec{y} \sim \Gamma(\overline{\nu}/2, \overline{\nu}\,\overline{s^2}/2)$.
  So,
  $$\mathsf{E}[\tau|\vec{y}] = 1/\overline{s^2}, \qquad Var[\tau|\vec{y}] = 2/(\overline{\nu}\,(\overline{s^2})^2).$$

Relationship to Frequentist Approach:

The PBE of β

$$\mathsf{E}[\beta|\vec{y}] = \overline{\beta} = \overline{c}(\underline{c}^{-1}\underline{\beta} + \widehat{\beta}\sum x_i^2).$$

It is a weighted average of the prior mean and the OLS estimator of β from frequentist statistics.

- $\underline{c}^{-1}$ reflects your confidence in the prior and should be chosen accordingly
- $\sum x_i^2$ reflects the degree of confidence that the data has in the OLS estimator $\widehat{\beta}$

Relationship to Frequentist Approach:

Recall also that

$$\overline{\nu}\overline{s^2} = \underline{\nu s^2} + \nu s^2 + \frac{(\widehat{\beta} - \underline{\beta})^2}{\underline{c} + \sum x_i^2}$$

and

$$s^2 = \frac{\sum (y_i - \widehat{\beta} x_i)^2}{n-1} = \frac{SSE}{n-1} = \frac{SSE}{\nu}.$$

So,

$$\underbrace{\overline{\nu}\overline{s^2}}_{\text{“posterior } SSE\text{”}} = \underbrace{\underline{\nu s^2}}_{\text{“prior } SSE\text{”}} + \underbrace{\nu s^2}_{SSE} + \frac{(\widehat{\beta} - \underline{\beta})^2}{\underline{c} + \sum x_i^2}$$

The final term reflects “conflict” between the prior and the data.

CHOOSING PRIOR HYPERPARAMETERS:

When choosing hyperparameters $\underline{\beta}$, $\underline{c}$, $\underline{\nu}$, and $\underline{s}^2$, it may be helpful to know that $\underline{\beta}$ is equivalent to the OLS estimate from an imaginary data set with

- $\underline{\nu} + 1$ observations
- imaginary $\sum x_i^2$ equal to $\underline{c}^{-1}$
- imaginary s^2 given by $\underline{s}^2$

The "imaginary" data set might even be previous data!

MODEL COMPARISON

Suppose you want to fit this overly simplistic linear model to describe the y_i but are not sure whether you want to use the x_i or a different set of explananatory variables.
Consider the two models:

$$M_1 \quad : \quad y_i = \beta_1 x_{1i} + \varepsilon_{1i}$$

$$M_2 \quad : \quad y_i = \beta_2 x_{2i} + \varepsilon_{2i}$$

Here, we assume

$$\varepsilon_{1i} \overset{iid}{\sim} N(0, \tau_1^{-1}) \qquad \text{and} \qquad \varepsilon_{2i} \overset{iid}{\sim} N(0, \tau_2^{-1})$$

are independent.

Model Comparison

- Priors for model j:

$$(\beta_j, \tau_j) \sim NG(\underline{\beta_j}, \underline{c_j}, \underline{\nu_j}/2, \underline{\nu}_j \underline{s_j}^2)$$

- $\Rightarrow$ posteriors for model j are

$$(\beta_j, \tau_j)|\vec{y} \sim NG(\overline{\beta_j}, \overline{c_j}, \overline{\nu_j}/2, \overline{\nu}_j \overline{s_j}^2)$$

- The posterior odds ratio is

$$PO_{12} := \frac{P(M_1|\vec{y})}{P(M_2|\vec{y})} = \frac{f(\vec{y}|M_1)}{f(\vec{y}|M_2)} \cdot \frac{P(M_1)}{P(M_2)}$$

Model Comparison

Can show that

$$f(\vec{y}|M_j) = a_j \left(\frac{\overline{c_j}}{\underline{c_j}} \right)^{1/2} \left(\overline{\nu_j}\, \overline{s_j^2} \right)^{\overline{\nu_j}/2}$$

where

$$a_j = \frac{\Gamma(\overline{\nu_j}/2) \cdot \left(\underline{\nu_j}\, \underline{s_j^2} \right)^{\underline{\nu_j}/2}}{\Gamma(\underline{\nu_j}/2) \cdot \pi^{n/2}}$$

MODEL COMPARISON

We can get the posterior model probabilities:

$$P(M_1|\vec{y}) = \frac{PO_{12}}{1 + PO_{12}}, \qquad P(M_2|\vec{y}) = \frac{1}{1 + PO_{12}}.$$

where

$$PO_{12} = \frac{a_1 \left(\frac{\overline{c_1}}{\underline{c_1}}\right)^{1/2} \left(\overline{\nu_1}\, \overline{s^2_1}\right)^{\overline{\nu_1}/2}}{a_2 \left(\frac{\overline{c_2}}{\underline{c_2}}\right)^{1/2} \left(\overline{\nu_2}\, \overline{s^2_2}\right)^{\overline{\nu_2}/2}} \cdot \frac{P(M_1)}{P(M_2)}$$

Model Comparison

$$PO_{12} = \frac{a_1 \left(\frac{\underline{c}_1}{\overline{c}_1} \right)^{1/2} \left(\overline{\nu}_1 \overline{s}_1^2 \right)^{\overline{\nu}_1/2}}{a_2 \left(\frac{\underline{c}_2}{\overline{c}_2} \right)^{1/2} \left(\overline{\nu}_2 \overline{s}_2^2 \right)^{\overline{\nu}_2/2}} \cdot \frac{P(M_1)}{P(M_2)}$$

- $\overline{\nu}_j \overline{s_j^2}$ contains the OLS SSE.
- A lower value indicates a better fit.
- So, the posterior odds ratio rewards models which fit the data better.

Model Comparison

$$PO_{12} = \frac{a_1 \left(\frac{\underline{c_1}}{\overline{c_1}}\right)^{1/2} \left(\overline{\nu_1}\,\overline{s_1^2}\right)^{\overline{\nu_1}/2}}{a_2 \left(\frac{\underline{c_2}}{\overline{c_2}}\right)^{1/2} \left(\overline{\nu_2}\,\overline{s_2^2}\right)^{\overline{\nu_2}/2}} \cdot \frac{P(M_1)}{P(M_2)}$$

- $\overline{\nu_j}\,\overline{s_j^2}$ contains a term like $(\widehat{\beta}_j - \underline{\beta_j})^2$
- So, the posterior odds ratio supports greater coherency between prior info and data info!