Bayesian Linear Regression

Yuki Shiraito

POLSCI 798 Advanced Topics in Quantitative Methodology
University of Michigan

Winter 2019

Linear Regression with Gaussian Errors

- Regression: Predicting conditional expectation of Y_i given X_i
- Linear regression: $\mathbb{E}[Y_i \mid \mathrm{X}_i] = \mathrm{X}_i^\top \beta$
- Ordinary linear regression:

$$Y_i = \mathrm{X}_i^\top \beta + \varepsilon_i,\ \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, N$$

$$\Leftrightarrow \mathrm{Y} \mid \beta, \sigma, \mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_N\right)$$

  - Unknown parameters: β and σ
  - $\mathbf{X}$ may be data, but inference is conditional on it
- Likelihood–multivariate Gaussian density:

$$p\left(\mathrm{Y} \,\middle|\, \beta, \sigma^2, \mathbf{X}\right)$$

$$= \frac{1}{\sqrt{\det\left(2\pi\sigma^2 \mathbf{I}\right)}} \exp\left(-\frac{1}{2}\left(\mathrm{Y} - \mathbf{X}\beta\right)^\top \left(\sigma^2 \mathbf{I}_N\right)^{-1} \left(\mathrm{Y} - \mathbf{X}\beta\right)\right)$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\mathrm{Y} - \mathbf{X}\beta\right)^\top \left(\mathrm{Y} - \mathbf{X}\beta\right)\right)$$

Variance Parameter

- Parameter σ^2 in the regression model:
  - *Not* the variance of Y_i, but the variance of ε_i
  - Cannot be estimated by the sample variance of Y_i
  - Need to be estimated simultaneously with β

$\rightsquigarrow$ Not assuming σ^2 is known, unlike randomized experiment

- Prior distribution $p\left(\beta, \sigma^2\right)$: Joint distribution on $\left(\beta, \sigma^2\right)$
- Conjugate prior on $\left(\beta, \sigma^2\right)$:

$$\sigma^2 \sim \text{Inv-}\chi^2\left(v_0, \sigma_0^2\right), \quad \beta \mid \sigma^2 \sim \mathcal{N}\left(\beta_0, \sigma^2 \Sigma_\beta\right)$$

  - Fully conjugate: Easier derivation of the posterior
  - Unclear interpretation: What is Σ_β?

- Semi-conjugate (conditionally conjugate) prior on $\left(\beta, \sigma^2\right)$:

$$\sigma^2 \sim \text{Inv-}\chi^2\left(v_0, \sigma_0^2\right), \quad \beta \sim \mathcal{N}\left(\beta_0, \Sigma_\beta\right)$$

  - Not fully conjugate: Trickier derivation
  - Clearer interpretation: Σ_β is the prior variance of β

Posterior with Conjugate Prior

- Prior density:

$$p\left(\beta,\sigma^2\right)=p\left(\sigma^2\right)p\left(\beta\,\middle|\,\sigma^2\right)\propto\frac{e^{-\frac{v_0\sigma_0^2}{2\sigma^2}}}{\left(\sigma^2\right)^{1+v/2}}\frac{e^{-\frac{1}{2\sigma^2}\left(\beta-\beta_0\right)^\top\Sigma_\beta^{-1}\left(\beta-\beta_0\right)}}{\left(\sigma^2\right)^{K/2}}$$

- Posterior density:

$$p\left(\beta,\sigma^2\,\middle|\,Y,\mathbf{X}\right)\propto\frac{e^{-\frac{v_0\sigma_0^2+Ns^2}{2\sigma^2}}}{\left(\sigma^2\right)^{1+(v_0+N)/2}}\frac{e^{-\frac{1}{2\sigma^2}\left(\beta-\widehat{\beta}_C\right)^\top\widehat{\Sigma}_C^{-1}\left(\beta-\widehat{\beta}_C\right)}}{\left(\sigma^2\right)^{K/2}},$$

$$\widehat{\Sigma}_C\equiv\left(\mathbf{X}^\top\mathbf{X}+\Sigma_\beta^{-1}\right)^{-1},\ \widehat{\beta}_C\equiv\widehat{\Sigma}_C\left(\mathbf{X}^\top Y+\Sigma_\beta^{-1}\beta_0\right),$$

$$s^2\equiv\frac{1}{N-K}\left(Y-\mathbf{X}\widehat{\beta}_C\right)^\top\left(Y-\mathbf{X}\widehat{\beta}_C\right)$$

- Factorization of the posterior:
  1. Conditional posterior of β given σ^2: $\beta\mid\sigma^2,Y,\mathbf{X}\sim\mathcal{N}\left(\widehat{\beta}_C,\sigma^2\widehat{\Sigma}_C\right)$
  2. Marginal posterior of σ^2: $\sigma^2\mid Y,\mathbf{X}\sim\text{Inv-}\chi^2\left(\widehat{v}_C,\widehat{\sigma}_C^2\right)$, where $\widehat{v}_C\equiv v_0+N$, $\widehat{\sigma}_C^2\equiv\left(v_0\sigma_0^2+Ns^2\right)/\widehat{v}_C$

Posterior with Semi-conjugate Prior

- Joint posterior density

$$p\left(\beta,\sigma^2 \,\middle|\, Y,\mathbf{X}\right) \propto p\left(\sigma^2\right) p\left(\beta\right) p\left(Y \,\middle|\, \beta,\sigma^2,\mathbf{X}\right)$$

$$\propto \frac{e^{-\frac{v_0\sigma_0^2}{2\sigma^2}}}{\left(\sigma^2\right)^{1+v_0/2}} e^{-\frac{1}{2}\left(\beta-\beta_0\right)^\top \Sigma_\beta^{-1}\left(\beta-\beta_0\right)} \frac{1}{\left(\sigma^2\right)^{N/2}} e^{-\frac{1}{2\sigma^2}\left(Y-\mathbf{X}\beta\right)^\top\left(Y-\mathbf{X}\beta\right)}$$

- Full conditional posterior:
  1. Conditional posterior of β given σ^2:

$$\beta \mid \sigma^2, Y, \mathbf{X} \sim \mathcal{N}\left(\widehat{\beta}_{SC}, \widehat{\Sigma}_{SC}\right)$$

$$\widehat{\Sigma}_{SC} \equiv \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}, \quad \widehat{\beta}_{SC} \equiv \widehat{\Sigma}_{SC}\left(\frac{1}{\sigma^2}\mathbf{X}^\top Y + \Sigma_\beta^{-1}\beta_0\right)$$

  2. Conditional posterior of σ^2 given β:

$$\sigma^2 \mid \beta, Y, \mathbf{X} \sim \text{Inv-}\chi^2\left(\widehat{v}_C, \widehat{\sigma}^2_{SC}\right)$$

$$S^2 \equiv \left(Y-\mathbf{X}\beta\right)^\top\left(Y-\mathbf{X}\beta\right), \quad \widehat{\sigma}^2_{SC} \equiv \left(v_0\sigma_0^2 + S^2\right)\Big/\widehat{v}_C$$

Gibbs Sampling

- Joint distributions:
  - Often difficult to summarize analytically
  - Need for Monte Carlo methods–simulations from a joint posterior
- Conjugate prior case:
  1. Draw σ^2 from the marginal posterior
     ```
     sigma2.samp <- rinvchisq(S, nu.c.hat, sigma2.c.hat)
     ```
  2. Draw β from the conditional posterior given σ^2
     ```
     for (s in 1:S) {
       beta.samp[s, ] <- mvrnorm(1, beta.c.hat, sigma2.samp[s] * Sigma.c.hat)
     }
     ```
- Semi-conjugate prior case:
  - Hard to draw a sample directly from the joint posterior
  - Need a method to draw from the joint using the full conditionals
- Gibbs sampling: Iterative sampling from full conditionals

  $s = 0$ Set an arbitrary initial value of σ^2

  $s = 1, 2, \ldots$ Repeat

  1. Draw $\beta^{(s)}$ from the conditional posterior given ${\sigma^2}^{(s-1)}$
  2. Draw ${\sigma^2}^{(s)}$ from the conditional posterior given $\beta^{(s)}$

Markov Chain

- Markov chain
  - Discrete-time stochastic process: $Z^{(1)}, Z^{(2)}, \ldots$
  - Markov property:
$$p\left(Z^{(s)} \middle| Z^{(1)}, \ldots, Z^{(s-1)}\right) = p\left(Z^{(s)} \middle| Z^{(s-1)}\right)$$
$$\Leftrightarrow \left(Z^{(1)}, \ldots, Z^{(s-1)}\right) \perp\!\!\!\perp \left(Z^{(s+1)}, \ldots\right) \middle| Z^{(s)}$$
  - E.g.: “Drunkard’s walk”
  - Gibbs sampling generates a Markov chain:
$$p\left(\beta^{(s)}, \sigma^{2\,(s)} \middle| \beta^{(1)}, \sigma^{2\,(1)}, \ldots, \beta^{(s-1)}, \sigma^{2\,(s-1)}, Y, \mathbf{X}\right)$$
$$= \underbrace{p\left(\sigma^{2\,(s)} \middle| \beta^{(s)}, Y, \mathbf{X}\right) p\left(\beta^{(s)} \middle| \sigma^{2\,(s-1)}, Y, \mathbf{X}\right)}_{=p\left(\beta^{(s)}, \sigma^{2\,(s)} \middle| \sigma^{2\,(s-1)}, Y, \mathbf{X}\right)}$$
- Stationary distribution, $p^*(Z)$
  - Invariance: $\forall z', p^*(Z = z') = \int_z p\left(Z^{(s+1)} = z' \middle| Z^{(s)} = z\right) p^*(Z = z)\, dz$
  - Under some conditions, the Markov chain:
    - Has a unique stationary distribution
    - Starts drawing samples from the stationary distribution (“converge”)

Stationary Distribution of the Gibbs Sampler

- Gibbs sampler for Bayesian linear regression:

$$\int\int p\left(\sigma^{2(s)}\middle|\beta^{(s)},\mathrm{Y},\mathbf{X}\right)p\left(\beta^{(s)}\middle|\sigma^{2(s-1)}=\sigma^2,\mathrm{Y},\mathbf{X}\right)\times p\left(\beta,\sigma^2\middle|\mathrm{Y},\mathbf{X}\right)d\beta d\sigma^2$$

$$=p\left(\sigma^{2(s)}\middle|\beta^{(s)},\mathrm{Y},\mathbf{X}\right)p\left(\beta^{(s)}\middle|\mathrm{Y},\mathbf{X}\right)=p\left(\beta^{(s)},\sigma^{2(s)}\middle|\mathrm{Y},\mathbf{X}\right)$$

- Generally, the Gibbs sampler for parameters $\theta\equiv(\theta_1,\dots,\theta_K)$
  - Iterations of alternating draws:
    - $\theta_1^{(s)}$ from $p\left(\theta_1\mid\theta_2^{(s-1)},\dots,\theta_K^{(s-1)},\text{Data}\right)$
    - $\theta_2^{(s)}$ from $p\left(\theta_2\mid\theta_1^{(s)},\theta_3^{(s-1)},\dots,\theta_K^{(s-1)},\text{Data}\right)$

      $\vdots$
    - $\theta_K^{(s)}$ from $p\left(\theta_K\mid\theta_1^{(s)},\dots,\theta_{K-1}^{(s)},\text{Data}\right)$
  - Stationary distribution:

$$\int p\left(\theta^{(s)}\mid\theta^{(s-1)}=\theta,\text{Data}\right)p\left(\theta\mid\text{Data}\right)d\theta=p\left(\theta^{(s)}\mid\text{Data}\right)$$

- The Gibbs sampler converges to the joint posterior distribution

Unequal Variances

- Pros of the Gibbs sampler:
  - Easy even for many parameters
  - Readily adapted to extended models with added parameters
- Linear regression with unequal variances (heteroscedasticity):

$$Y_i = X_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{j[i]}\right)$$

  - J groups of observations, and obs. i belongs to group $j[i]$
  - Group-specific variance, $\sigma^2_{j[i]}$
  - Randomized experiment as a special case
- Joint posterior density of $\beta, \sigma^2_1, \ldots, \sigma^2_J$:

$$p\left(\beta, \sigma^2 \,\middle|\, \mathrm{Y}, \mathbf{X}\right) \propto \prod_{j=1}^{J} \frac{e^{-\frac{v_0 \sigma_0^2}{2\sigma_j^2}}}{\left(\sigma_j^2\right)^{1+v_0/2}} e^{-\frac{1}{2}\left(\beta-\beta_0\right)^\top \Sigma_\beta^{-1}\left(\beta-\beta_0\right)} \times \frac{1}{\left(\sigma_j^2\right)^{N_j/2}} e^{-\frac{1}{2\sigma_j^2}\left(\mathrm{Y}_j-\mathbf{X}_j\beta\right)^\top\left(\mathrm{Y}_j-\mathbf{X}_j\beta\right)}$$

where $\mathrm{Y}_i, \mathbf{X}_j$ are the data of group j

Modifying the Gibbs Sampler

- Conditional posterior density of:

  σ_j^2 – Equivalent to using the data of group j only

$$\sigma_j^2 \mid \beta, \mathrm{Y}, \mathbf{X} \sim \text{Inv-}\chi^2\left(\widehat{v}_{UV,j}, \widehat{\sigma}^2_{UV,j}\right)$$

$$\widehat{v}_{UV,j} \equiv v_0 + N_j,\ \ S_j^2 \equiv \left(\mathrm{Y}_j - \mathbf{X}_j\beta\right)^\top \left(\mathrm{Y}_j - \mathbf{X}_j\beta\right),\ \widehat{\sigma}^2_{UV,j} \equiv \frac{\left(v_0\sigma_0^2 + S_j^2\right)}{\widehat{v}_{UV,j}}$$

  β – Equivalent to adding the data of each group sequentially

$$\beta \mid \sigma^2_{1:J}, \mathrm{Y}, \mathbf{X} \sim \mathcal{N}\left(\widehat{\beta}_{UV}, \widehat{\Sigma}_{UV}\right)$$

$$\widehat{\Sigma}_{UV} \equiv \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2}\mathbf{X}_j^\top\mathbf{X}_j + \Sigma_\beta^{-1}\right)^{-1},\ \widehat{\beta}_{UV} \equiv \widehat{\Sigma}_{UV}\left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2}\mathbf{X}_j^\top\mathrm{Y}_j + \Sigma_\beta^{-1}\beta_0\right)$$

- Gibbs sampler:
  1. Draw $\beta^{(s)}$ given $\sigma_1^{2\,(s-1)}, \ldots, \sigma_J^{2\,(s-1)}$
  2. Draw $\sigma_1^{2\,(s)}, \ldots, \sigma_J^{2\,(s)}$ independently conditional on $\beta^{(s)}$
- Conditioning on all the other parameters

  $\rightsquigarrow$ Can ignore changes of the other parameters' sampling

Prior as Additional "Data Points"

- Prior:
  - Extra source of information
  - "Addtional data points" (c.f. Beta-Binomial model)

- "Additional data points" for linear regression:

  β – K observations with $Y = \beta_0, \mathbf{X} = \mathbf{I}_K$, and known variance Σ_β

  E.g., If $\Sigma_\beta = \sigma_\beta^2 \mathbf{I}_K$,

$$\widehat{\Sigma}_{SC} = \left(\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X} + \frac{1}{\sigma_\beta^2} \mathbf{I}_K^\top \mathbf{I}_K \right)^{-1}, \; \widehat{\beta}_{SC} = \widehat{\Sigma}_{SC} \left(\frac{1}{\sigma^2} \mathbf{X}^\top Y + \frac{1}{\sigma_\beta^2} \mathbf{I}_K^\top \beta_0 \right)$$

  σ^2 – v_0 observations with sample variance σ_0^2

$$\widehat{v}_C = \underbrace{v_0}_{\text{prior "}N\text{"}} + N,$$

$$\widehat{v}_C \widehat{\sigma}_{SC}^2 = \underbrace{v_0 \sigma_0^2}_{\text{prior "}N\text{"} \times \text{prior sample variance}} + \underbrace{S^2}_{N \times \text{mean squared residuals}}$$

  - Same principle applies to the conjugate case as well

Noninformative Prior and MLE

- Noninformative prior: What does "noninformative" mean?
  - "*Zero* additional data points":

$$p\left(\sigma^2\right)p\left(\beta\right)\propto\frac{e^{-\frac{\overbrace{v_0}^{\to 0}\sigma_0^2}{2\sigma^2}}}{\underbrace{\left(\sigma^2\right)^{1+\underbrace{v_0}_{\to 0}/2}}_{\to 1/\sigma^2}}\underbrace{e^{-\frac{1}{2}\left(\beta-\beta_0\right)^\top\Sigma_\beta^{-1}\left(\beta-\beta_0\right)}}_{\to 1}$$

$$\Rightarrow p\left(\sigma^2,\beta\right)\propto\frac{1}{\sigma^2}\Leftrightarrow p\left(\log\sigma,\beta\right)\propto 1$$

  - *Not necessarily uniform*
  - *Not necessarily invariant to transformation*
  - Improper for unbounded parameters: $\int_0^\infty\int_{-\infty}^\infty\frac{1}{\sigma^2}d\beta d\sigma^2=\infty$
  - Leads to proper posterior in linear regression, but not always
- Often recommended "weakly informative prior":
  - $\sigma\sim\text{Unif}(0,r)$ for large r
  - Allowing heavier tail while preventing diverge
  - Posterior mode $\to$ MLE (c.f. Beta-Binomial model)

Posterior Predictive Distributions

- Purpose of regression analysis: Prediction
  - Fit the model to a data set $(\mathbf{X}, \mathrm{Y})$
  - If observe a new data point X_{new}, what is Y_{new} given the model?
- Posterior predictive distribution of Y_{new}:

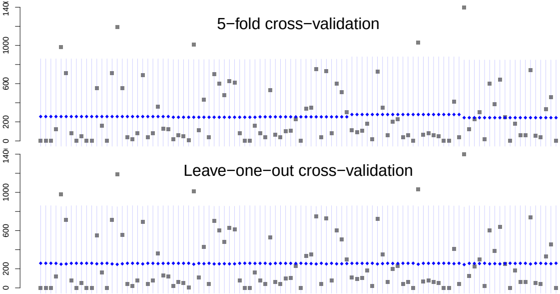
$$p\left(Y_{\text{new}} \,|\, \mathrm{X}_{\text{new}}, \mathbf{X}, \mathrm{Y}\right) = \int\int p\left(Y_{\text{new}}, \beta, \sigma^2 \,\middle|\, \mathrm{X}_{\text{new}}, \mathbf{X}, \mathrm{Y}\right) d\beta d\sigma^2$$

$$= \int\int \underbrace{p\left(Y_{\text{new}} \,\middle|\, \mathrm{X}_{\text{new}}, \beta, \sigma^2\right)}_{\text{Model}} \underbrace{p\left(\beta, \sigma^2 \,\middle|\, \mathbf{X}, \mathrm{Y}\right)}_{\text{Posterior}} d\beta d\sigma^2$$

- Computation via Monte Carlo:
  1. Draw of $\left(\beta^{(s)}, \sigma^{2\,(s)}\right)$ from the posterior
  2. Draw $Y_{\text{new}}^{(s)}$ from $\mathcal{N}\left(\mathrm{X}_{\text{new}}^{\top}\beta^{(s)}, \sigma^{2\,(s)}\right)$
- Goodness of fit: Out-of-sample prediction
  - How good $p\left(Y_{\text{new}} \,|\, \mathrm{X}_{\text{new}}, \mathbf{X}, \mathrm{Y}\right)$ is as a prediction of Y_{new}
  - Need new data, but data collection is usually one-shot
    $\rightsquigarrow$ cross-validation

Cross-validation

- Cross-validation:
  - Approximation for out-of-sample prediction
  - K-fold cross-validation
    1. Split the data into K subsets
    2. Hold a subset out as a test set
    3. Fit the model to the remaining $K-1$ subsets
    4. Check the prediction on the test set
    5. Repeat K times so each subset is held-out once
  - Leave-one-out cross-validation (LOO-CV): $K = N$



Summary

- Linear regression with Gaussian errors
  - Building block of many other models
  - Parameters: β and σ^2
  - Conjugate and semi-conjugate priors
    - Conjugate: Prior dependence, mathematical tractability
    - Semi-conjugate: Prior independence, no explicit joint posterior
- Gibbs sampler
  - Alternating draws from full conditional posteriors
  - Distribution of draws converges to the joint posterior
  - Semi-conjugate model, extensions
- Posterior predictive distributions and cross-validation
- Readings for review:
  1. Noninformative prior: **BDA3** Sections 2.8–9
  2. Bayesian linear regression: **BDA3** Chapters 3 and 14
  3. Gibbs sampler: **BDA3** Sections 11.0-1
  4. (Optional) Asymptotic approximation: **BDA3** Chapter 4