Notes on Bayesian Linear Regression

Max Leung

April 19, 2024

1 Joint Posterior Distribution

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) &= p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X})p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}) \\
&= p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}) \underbrace{p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})}_{\int p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})d\sigma^2}
\end{aligned}
$$

1.1 Normally Distributed and Homoscedastic y with Non-informative Prior

1.1.1 Marginal Posterior Distribution

$$
\begin{aligned}
p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) &\propto L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X})\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{X}) \\
&= N(\mathbb{E}(\boldsymbol{y}|\boldsymbol{X}), Var(\boldsymbol{y}|\boldsymbol{X})) \cdot C \\
&\propto N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \\
&= \frac{1}{(2\pi)^{n/2}det(\sigma^2\boldsymbol{I})^{1/2}} exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}det(\boldsymbol{I})^{1/2}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}+\boldsymbol{X}\widehat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}((\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})-\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}))'((\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})-\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}))) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\varepsilon}}'-(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}')(\widehat{\boldsymbol{\varepsilon}}-\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}))) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}-\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})-(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\widehat{\boldsymbol{\varepsilon}}+(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}))) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}+(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}))) && \text{since } \boldsymbol{X}'\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{0} \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{1}{2\sigma^2}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}) exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{n-k}{2\sigma^2}\widehat{\sigma}^2) exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{n-k}{2\sigma^2}\widehat{\sigma}^2) \frac{(2\pi)^{k/2}det(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})^{1/2}}{(2\pi)^{k/2}det(\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})^{1/2}} exp(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})) \\
&= \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp(-\frac{n-k}{2\sigma^2}\widehat{\sigma}^2)(2\pi)^{k/2}(\sigma^2)^{k/2}det((\boldsymbol{X}'\boldsymbol{X})^{-1})^{1/2} N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&= \frac{det((\boldsymbol{X}'\boldsymbol{X})^{-1})^{1/2}}{(2\pi)^{(n-k)/2}} (\sigma^2)^{-(n-k)/2} exp(-\frac{n-k}{2\sigma^2}\widehat{\sigma}^2) N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&= \frac{det((\boldsymbol{X}'\boldsymbol{X})^{-1})^{1/2}}{(2\pi)^{(n-k)/2}C_1} C_1(\sigma^{-2})^{(n-k)/2+1} exp(-\frac{(n-k)\widehat{\sigma}^2}{2}/\sigma^2) N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&\propto IG(\sigma^2|\frac{n-k}{2}, \frac{(n-k)\widehat{\sigma}^2}{2}) N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) && \text{due to } IG(z|a,b) = \tfrac{b^a}{\Gamma(a)}(1/z)^{a+1}exp(-b/z) \\
&= N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})\chi^{-2}(\sigma^2|n-k, \widehat{\sigma}^2) && \text{due to } IG(z|\nu/2, \nu\tau^2/2) = \chi^{-2}(z|\nu, \tau^2)
\end{aligned}
$$

Thus, $p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) = N(\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ and $p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}) = \chi^{-2}(\sigma^2|n-k, \widehat{\sigma}^2)$.

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) &= \int p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) d\sigma^2 \\
&= \int p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}) p(\sigma^2|\boldsymbol{y}, \boldsymbol{X}) d\sigma^2 \\
&= \int N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \chi^{-2}(\sigma^2|n-k, \widehat{\sigma}^2) d\sigma^2 \\
&= t(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, (\boldsymbol{X}'\boldsymbol{X})^{-1}, n-k)
\end{aligned}$$

1.1.2 Algorithm for sampling from joint posterior distribution

Given the closed form solutions of the marginal posterior distribution, we can sample from joint posterior distribution without using Metropolis-Hastings algorithm,

Step 1 - compute $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}/(n-k) = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/(n-k)$
Step 2 - draw σ^2 from $\chi^{-2}(\sigma^2|n-k, \widehat{\sigma}^2)$
Step 3 - draw $\boldsymbol{\beta}$ from $N(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$

In order to speed up the computation, we can apply QR decomposition of $\boldsymbol{X}$ to compute both $(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1'}$ and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ which is the numerical solution of $\boldsymbol{R}\widehat{\boldsymbol{\beta}} = \boldsymbol{Q}'\boldsymbol{y}$. Read Gelman, et al. (2013, p.356) for the details.

The following R code performs the above algorithm.

```
library(LearnBayes)
set.seed(15)

sim_beta <- list()
sim_sigma_square <- list()
# lm() performs QR decomposition under the hood
ols <- lm(y ~ x1 + x2 + x3, data = d, method = "qr")
n_minus_k <- ols$df.residual
beta_hat <- ols$coefficients
sigma_square_hat <- sum(ols$residuals^2, na.rm = TRUE) / n_minus_k
XX_inverse <- vcov(ols) / sigma_square_hat

for (i in seq_len(2000)) {
  sim_sigma_square[[i]] <- LearnBayes::rigamma(1, n_minus_k / 2, n_minus_k * sigma_square_hat / 2)
  sim_beta[[i]] <- rmnorm(1, beta_hat, sim_sigma_square[[i]] * XX_inverse)
}

sim_beta_m <- do.call(rbind, sim_beta)
sim_sigma_square_m <- do.call(rbind, sim_sigma_square)

colMeans(sim_beta_m, na.rm = TRUE)
apply(sim_beta_m, 2, quantile, c(0.25, 0.5, 0.75), na.rm = TRUE)

mean(sim_sigma_square_m, na.rm = TRUE)
quantile(sim_sigma_square_m, c(0.25, 0.5, 0.75), na.rm = TRUE)
```

blinreg function from *LearnBayes* package does the same thing.

```
library(LearnBayes)
set.seed(15)

linear_reg_u_prior <-
  blinreg(
    y = d[["y"]],
    X = as.matrix(cbind(1, d[c("x1", "x2", "x3")])),
    m = 1000,
    prior = NULL
  )

colMeans(linear_reg_u_prior$beta, na.rm = TRUE)
apply(linear_reg_u_prior$beta, 2, quantile, c(0.25, 0.5, 0.75), na.rm = TRUE)

mean(linear_reg_u_prior$sigma, na.rm = TRUE)
quantile(linear_reg_u_prior$sigma, c(0.25, 0.5, 0.75), na.rm = TRUE)
```

The package *rstanarm* in R offers another easy-to-use function to draw samples from joint posterior distribution with Hamiltonian Monte Carlo (HMC) algorithm and QR decomposition.

```
library(rstanarm)

stan_glm(
  y ~ x1 + x2 + x3,
  family = gaussian(),
  prior_intercept = NULL,
  prior = NULL,
  prior_aux = NULL,
  algorithm = "sampling",
  QR = TRUE,
  data = d
)
```

1.2 t Distributed y

It is the bayesian estimation of robust student t regression.

1.2.1 Joint Posterior Distribution

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\Omega}, \nu | \boldsymbol{y}, \boldsymbol{X}) &\propto L(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{\Omega}, \nu, \boldsymbol{X}) \pi(\boldsymbol{\beta}, \boldsymbol{\Omega}, \nu | \boldsymbol{X}) \\ &= t(\boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{\Omega}, \nu, \boldsymbol{X}) \cdot \pi(\boldsymbol{\beta}, \boldsymbol{\Omega}, \nu | \boldsymbol{X}) \\ &= \int N(\boldsymbol{y} | \boldsymbol{\beta}, z \cdot \boldsymbol{\Omega}, \boldsymbol{X}) \chi^{-2}(z | \nu, 1) dz \cdot \pi(\boldsymbol{\beta}, \boldsymbol{\Omega}, \nu | \boldsymbol{X}) \qquad \text{where } \boldsymbol{\Omega} = diag(\sigma^2) \end{aligned}$$

The joint posterior distribution can be sampled by using Metropolis-Hastings algorithm (including Gibbs Sampler). The following R code demonstrates an example where $\pi(\boldsymbol{\beta}, \boldsymbol{\Omega}, \nu | \boldsymbol{X}) = \prod_j N(\beta_j | 0, 10^2) \cdot Gamma(\sigma^2 | 2, 0.1) \cdot Gamma(\nu | 2, 0.1)$.

```
library(runjags)
set.seed(15)

model_string =
  "
    model {
      for (i in 1:N) {
        # variance of y_i = z_i sigma^2
        # precision of y_i = 1 / (z_i sigma^2) = (1 / z_i) / sigma^2 = phi_i / sigma^2
        y[i] ~ dnorm(mu[i], phi[i] / sigma_square)
        mu[i] <- b[1] + b[2]*x1[i] + b[3]*x2[i] + b[4]*x3[i]
        phi[i] ~ dgamma(nu / 2, nu / 2)
      }
      # prior
      for (j in 1:4) {
        b[j] ~ dnorm(0, 1 / 100)
      }
      sigma_square ~ dgamma(2, 0.1)
      nu ~ dgamma(2, 0.1)
    }
  "

model_data <-
  list(
    "y" = d[["y"]],
    "N" = length(d[["y"]]),
    "x1" = d[["x1"]],
    "x2" = d[["x2"]],
    "x3" = d[["x3"]]
  )

t_reg_jags <-
  run.jags(
    model = model_string,
    n.chains = 1,
    data = model_data,
    monitor = c("b", "sigma_square", "nu"),
    adapt = 1000,
    burnin = 5000,
    sample = 5000
  )

print(t_reg_jags)
plot(t_reg_jags, var = "b")
```

```
plot(t_reg_jags, var = "sigma_square")
plot(t_reg_jags, var = "nu")
```

The *brms* package in R offers simpler code.

```
library(brms)

brm(
  data = d,
  family = student,
  y ~ 1 + x1 + x2 + x3,
  prior = c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b),
    prior(gamma(2, 0.1), class = nu),
    prior(gamma(2, 0.1),  class = sigma)
  ),
  seed = 15
)
```

2 Reference

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and Other Stories. Cambridge: Cambridge University Press.