An Introduction to Bayesian Statistics - I

An-Ting Jhuang

UnitedHealth Group R&D

ajhuang@savvysherpa.com

August 16, 2019

Overview

1. Big idea behind Bayesian statistics
   - Bayes' theorem
   - Terminology explanation
   - Bayesian vs frequentist
2. Prior types
   - By reasoning
   - By mathematical property
3. Three examples in R
   - Coin flip example
   - Logistic regression example
   - Linear regression example

Big Idea behind Bayesian Statistics

Figure 1: Thomas Bayes, 1701-1761

- Treat probability as a **degree of belief**

Big Idea behind Bayesian Statistics

Figure 1: Thomas Bayes, 1701-1761

- Treat probability as a **degree of belief**
- Continually update our prior beliefs about events as new evidence is presented

Big Idea behind Bayesian Statistics

Figure 1: Thomas Bayes, 1701-1761

- Treat probability as a **degree of belief**
- Continually update our prior beliefs about events as new evidence is presented
- Probabilistic reasoning leads to probabilistic results

Bayes Reasoning

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Reasoning

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior Distribution

$$P(A|B) \propto P(B|A) \times P(A)$$

$$P(\text{parameter}|\text{data}) \propto P(\text{data}|\text{parameter}) \times P(\text{parameter})$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Terminology

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- **Prior**: the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account

Terminology

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- **Prior**: the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account
- **Likelihood**: the probability distribution of evidence given parameters

Terminology

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

- **Prior**: the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account
- **Likelihood**: the probability distribution of evidence given parameters
- **Posterior**: the probability distribution of parameters given evidence

Bayesian vs Frequentist

Table 1: Comparison between frequentist and Bayesian

Characteristic	Frequentist	Bayesian
Probability	limiting relative frequency	degree of belief
Parameter	fixed constant	random variable
Probability statement	procedure	parameter

Prior Types

- By reasoning,
  - Expert prior: a prior presenting expert knowledge
- By mathematical property,

Prior Types

- By reasoning,
  - Expert prior: a prior presenting expert knowledge
  - Uninformative prior: a prior with big variance
- By mathematical property,

Prior Types

- By reasoning,
  - Expert prior: a prior presenting expert knowledge
  - Uninformative prior: a prior with big variance
  - Objective prior: a prior in the absence of prior information
- By mathematical property,

Prior Types

- By reasoning,
  - Expert prior: a prior presenting expert knowledge
  - Uninformative prior: a prior with big variance
  - Objective prior: a prior in the absence of prior information
- By mathematical property,
  - Conjugate prior: leads to a posterior from the same parametric family as the prior

Prior Types

- By reasoning,
  - Expert prior: a prior presenting expert knowledge
  - Uninformative prior: a prior with big variance
  - Objective prior: a prior in the absence of prior information
- By mathematical property,
  - Conjugate prior: leads to a posterior from the same parametric family as the prior
  - Non-conjugate prior: does not result in a posterior from the same parametric family as the prior

Coin Flip Example

- Let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses
- What's the estimate of θ?

Coin Flip Example

- Let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses
- What's the estimate of θ?
- Frequentist: $\hat{\theta}_{\text{ML}} = y/n$

Coin Flip Example

- Let θ be the probability of heads, n be the number of tosses, and y be the number of heads in n tosses
- What's the estimate of θ?
- Frequentist: $\hat{\theta}_{\text{ML}} = y/n$
- Bayesian:
  (1) choose a prior of θ
  (2) calculate posterior distribution
  (3) $\hat{\theta}_{\text{Bayes}} =$ posterior mean

Coin Flip Example - Conjugate Prior

- Consider a conjugate prior $\theta \sim \text{Beta}(\alpha, \beta)$ and the likelihood $y|\theta \sim \text{Bin}(n, \theta)$, then

$$
\begin{aligned}
\theta|y &\propto \text{prior} \times \text{likelihood} \\
&\propto \text{Beta}(\alpha, \beta) \times \text{Bin}(n, \theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{y}(1-\theta)^{n-y} \\
&= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.
\end{aligned}
$$

Coin Flip Example - Conjugate Prior

- Consider a conjugate prior $\theta \sim \text{Beta}(\alpha, \beta)$ and the likelihood $y|\theta \sim \text{Bin}(n, \theta)$, then

$$\begin{aligned}
\theta|y &\propto \text{prior} \times \text{likelihood} \\
&\propto \text{Beta}(\alpha, \beta) \times \text{Bin}(n, \theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{y}(1-\theta)^{n-y} \\
&= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.
\end{aligned}$$

- The posterior distribution $\theta|y \sim \text{Beta}(\alpha + y, \beta + n - y)$.

Coin Flip Example - Conjugate Prior

- Consider a conjugate prior $\theta \sim \text{Beta}(\alpha, \beta)$ and the likelihood $y|\theta \sim \text{Bin}(n, \theta)$, then

$$\begin{aligned}
\theta|y &\propto \text{prior} \times \text{likelihood} \\
&\propto \text{Beta}(\alpha, \beta) \times \text{Bin}(n, \theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^{y}(1-\theta)^{n-y} \\
&= \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}.
\end{aligned}$$

- The posterior distribution $\theta|y \sim \text{Beta}(\alpha + y, \beta + n - y)$.
- The Bayes estimator is the posterior mean, $\hat{\theta}_{\text{Bayes}} = \frac{\alpha+y}{\alpha+\beta+n}$.

Coin Flip Example - Conjugate Prior

- Set up

```
##Coin flip
set.seed(98712)
n        <- n_post <- 10^5
y        <- rbinom(1,n,.48)
alpha    <- 2
beta     <- 2
```

- Specify prior and likelihood

```
set.seed(3878)
th_pror   <- rbeta(n_post,alpha,beta)
plot(density(th_pror),xlab=expression(theta[prior]),main="")

y_like    <- rbinom(n,n,.48)
plot(density(y_like),xlab="y, number of heads",main="")
```

Coin Flip Example - Conjugate Prior

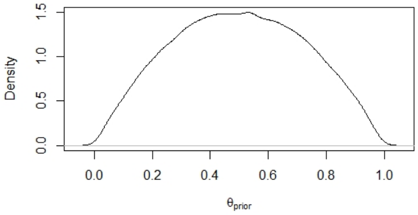

Figure 2: Prior density of θ

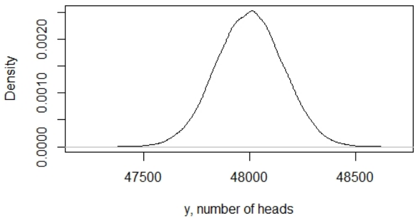

Figure 3: Likelihood of $y|\theta$

Coin Flip Example - Conjugate Prior

- Draw a posterior sample

```
th_post  <- rbeta(n_post,alpha+y,beta+n-y)
plot(density(th_post),xlab=expression(theta[post]),main="")
abline(v=y/n,col="darkgreen",lty=2,lwd=2)
abline(v=(alpha+y)/(alpha+beta+n),col="blue",lwd=2)
legend(.471,250,c("ML","Bayes"),col=c("darkgreen","blue"),
       pch=rep(19,2),bty="n")
legend(.478,250,"ML and Bayes estimates \n almost the same",bty="n")
```

Coin Flip Example - Conjugate Prior

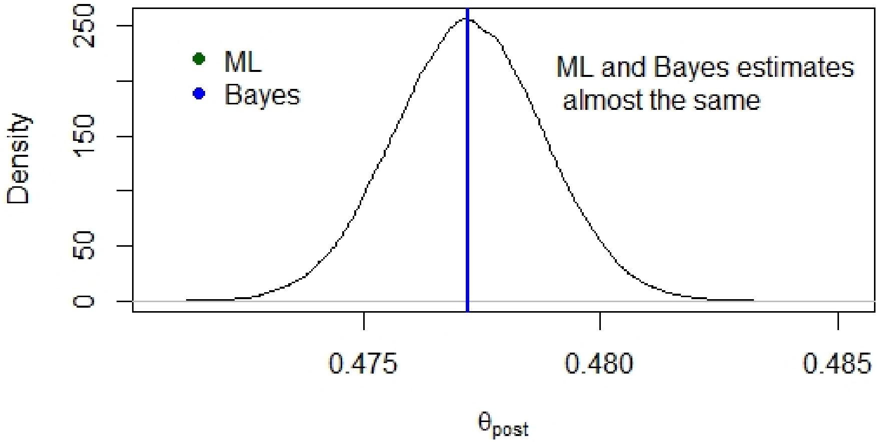

Figure 4: Posterior density of θ

Moving Away From a Conjugate Prior

- What do we do without a conjugate prior?

Moving Away From a Conjugate Prior

- What do we do without a conjugate prior?
- Posterior probably won't be a recognizable distribution, so you will need to work a little harder to conduct inference.

Moving Away From a Conjugate Prior

- What do we do without a conjugate prior?
- Posterior probably won't be a recognizable distribution, so you will need to work a little harder to conduct inference.
- For example, in a logistic regression model $\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 x$, what if we have a normal prior $N(0, \sigma^2)$ for β_0 and β_1?

Moving Away From a Conjugate Prior

- What do we do without a conjugate prior?
- Posterior probably won't be a recognizable distribution, so you will need to work a little harder to conduct inference.
- For example, in a logistic regression model $\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 x$, what if we have a normal prior $N(0, \sigma^2)$ for β_0 and β_1?
- The posterior $P(\beta_k|y) \propto p^y(1-p)^{(n-y)} \times exp^{-\frac{1}{2\sigma^2}\beta_k^2}$ doesn't lead to a recognizable distribution.

Moving Away From a Conjugate Prior

- What do we do without a conjugate prior?
- Posterior probably won't be a recognizable distribution, so you will need to work a little harder to conduct inference.
- For example, in a logistic regression model $\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 x$, what if we have a normal prior $N(0, \sigma^2)$ for β_0 and β_1?
- The posterior $P(\beta_k|y) \propto p^y(1-p)^{(n-y)} \times exp^{-\frac{1}{2\sigma^2}\beta_k^2}$ doesn't lead to a recognizable distribution.
- How can we conduct inference with this beast of a posterior distribution?

Moving Away From a Conjugate Prior

Basic Bayesian Steps

1. Select a model and priors
2. Approximate the posterior via Markov chain Monte Carlo
3. Check the posterior approximation (e.g. sufficient samples)
4. Use the MCMC samples to conduct inference

You can either code it yourself or use one of the MANY packages on CRAN.

Approximating the Posterior via MCMC

- Approximate the posterior by using Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.

Approximating the Posterior via MCMC

- Approximate the posterior by using Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.
- How does MCMC sampling generally work?
  1. Select a starting value for each parameter

Approximating the Posterior via MCMC

- Approximate the posterior by using Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.
- How does MCMC sampling generally work?
  1. Select a starting value for each parameter
  2. Iterate between the following two steps:
     1. Propose new values based on the current parameter values
     2. Move to the proposed values with some probability, or stay at the current position with the complementary probability

Approximating the Posterior via MCMC

- Approximate the posterior by using Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.
- How does MCMC sampling generally work?
  1. Select a starting value for each parameter
  2. Iterate between the following two steps:
     1. Propose new values based on the current parameter values
     2. Move to the proposed values with some probability, or stay at the current position with the complementary probability
- The exact method of selecting proposed values and calculating the probability of moving depends on the exact MCMC sampler.

Approximating the Posterior via MCMC

- Some packages (such as `mcmc`) focus on the MCMC (independent of the model/context) and are therefore **more general**.
- `mcmc` simulates using a user-inputted log unnormalized posterior density.

Approximating the Posterior via MCMC

- Some packages (such as `mcmc`) focus on the MCMC (independent of the model/context) and are therefore **more general**.
- `mcmc` simulates using a user-inputted log unnormalized posterior density.
- Some packages (such as `MCMCpack`) contain functions to perform **specific methods of Bayesian inference**.
- `MCMCpack` does MCMC in the context of specific statistical models.

Logistic Regression Example - Non-conjugate Prior

- Set up & construct the log posterior density

```r
library(mcmc)
data(logit)
out <- glm(y~x1,data=logit,family=binomial,x=TRUE)

lupost_factory <- function(x,y)function(beta){
   eta  <- as.numeric(x%*%beta)
   logp <- ifelse(eta<0,eta-log1p(exp(eta)),-log1p(exp(-eta)))
   logq <- ifelse(eta<0,-log1p(exp(eta)),-eta-log1p(exp(-eta)))
   logl <- sum(logp[y==1])+sum(logq[y==0])
   return(logl-sum(beta^2)/8)
   }
lupost <- lupost_factory(out$x,out$y)
```

- Draw a posterior sample

```r
set.seed(317)
beta.init <- as.numeric(coefficients(out))
out <- metrop(lupost,beta.init,1e3)
names(out)
out$accept
```

Logistic Regression Example - Non-conjugate Prior

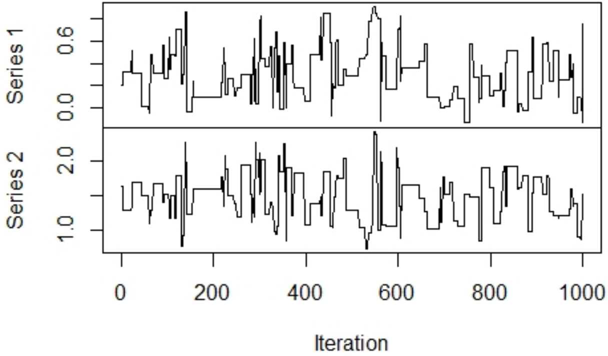

Figure 5: Trace plot of $\hat{\beta}_0$ and $\hat{\beta}_1$

Linear Regression Example

- **Bikeshare dataset**: contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Linear Regression Example

- **Bikeshare dataset**: contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.
- **Objective**: investigate if feel like temperature is significantly related to number of registered riders.

Linear Regression Example

- **Bikeshare dataset**: contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.
- **Objective**: investigate if feel like temperature is significantly related to number of registered riders.

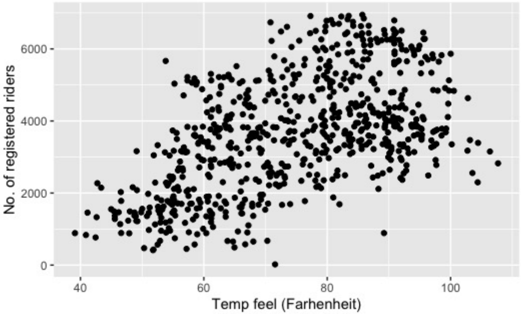

Figure 6: Scatter plot of feel like temperature and number of registered riders

Linear Regression Example

- Let Y_i be the number of registered riders and x_i be the feels like temperature (Fahrenheit) on date $i = 1, ..., n$, then the linear regression model is
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Linear Regression Example

- Let Y_i be the number of registered riders and x_i be the feels like temperature (Fahrenheit) on date $i = 1, ..., n$, then the linear regression model is
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$
- Random error: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Linear Regression Example

- Let Y_i be the number of registered riders and x_i be the feels like temperature (Fahrenheit) on date $i = 1, ..., n$, then the linear regression model is
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$
- Random error: $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$
- Prior specification: $\beta_0, \beta_1 \overset{indep}{\sim} N(\mu_{\beta_k}, \sigma^2_{\beta_k}), \sigma^2 \sim \text{InvGamma}(a, b)$

Linear Regression Example

- Fit the model in two ways:

```
#fit a regression model
freq.fit  <- lm(riders_registered ~ temp_feel, data=bikes)

bayes.fit <- MCMCregress(riders_registered~temp_feel,
                         b0=freq.fit$coefficients,B0=.01,
                         c0=.1,d0=.1,
                         beta.start=freq.fit$coefficients,
                         data=bikes,
                         burnin=10^3,mcmc=10^5)
```

Linear Regression Example

- Very close regression lines: $\hat{y} = -667.9 + 57.9x$

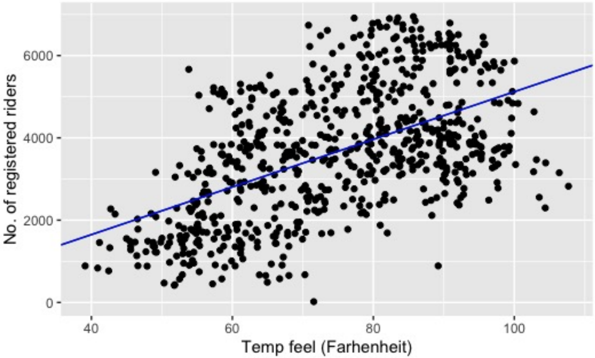

Figure 7: Regression lines and scatter plot of feel like temperature and number of registered riders

Linear Regression Example

- Is feel like temperature significantly related to number of registered riders?

Linear Regression Example

- Is feel like temperature significantly related to number of registered riders?
- Hypothesis test: $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$

Linear Regression Example

- Is feel like temperature significantly related to number of registered riders?
- Hypothesis test: $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -667.916    251.608  -2.655  0.00811 **
temp_feel     57.892      3.306  17.514  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1310 on 729 degrees of freedom
Multiple R-squared:  0.2961,	Adjusted R-squared:  0.2952
F-statistic: 306.7 on 1 and 729 DF,  p-value: < 2.2e-16
```

Figure 8: Coefficient summary of ordinary least square estimates

Linear Regression Example

```
Iterations = 1001:101000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1e+05

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                  Mean        SD  Naive SE Time-series SE
(Intercept)    -667.90 9.975e+00 3.154e-02      3.154e-02
temp_feel        57.89 6.474e-01 2.047e-03      2.047e-03
sigma2      1718089.06 9.051e+04 2.862e+02      2.862e+02
```

Figure 9: Coefficient summary of Bayesian estimates

Linear Regression Example

```
Iterations = 1001:101000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1e+05

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                  Mean        SD  Naive SE Time-series SE
(Intercept)    -667.90 9.975e+00 3.154e-02      3.154e-02
temp_feel        57.89 6.474e-01 2.047e-03      2.047e-03
sigma2      1718089.06 9.051e+04 2.862e+02      2.862e+02
```

Figure 9: Coefficient summary of Bayesian estimates

```
2. Quantiles for each variable:

                  2.5%        25%        50%        75%      97.5%
(Intercept)    -687.45    -674.65    -667.88    -661.21    -648.29
temp_feel        56.62      57.45      57.89      58.33      59.17
sigma2      1550133.89 1655338.50 1714929.37 1777374.65 1904058.94
```

Figure 10: Quantile summary of Bayesian estimates

Linear Regression Example

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\mathrm{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.
- In Bayesian statistics, a 95% credible set of β_1 is (56.6, 59.2).

Linear Regression Example

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\mathrm{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.
- In Bayesian statistics, a 95% credible set of β_1 is (56.6, 59.2).
- Different interpretations:

	Term	Meaning
Frequentist	confidence interval	95% certain $\beta_1 \in (51.4, 64.4)$
Bayesian	credible set	$P(56.6 \leq \beta_1 \leq 59.2) = 0.95$

Linear Regression Example

- In frequentist method, a 95% confidence interval of β_1 is $\hat{\beta}_1 \pm t_{0.025,729}\mathrm{SE}(\hat{\beta}_1) \approx (51.4, 64.4)$.
- In Bayesian statistics, a 95% credible set of β_1 is (56.6, 59.2).
- Different interpretations:

	Term	Meaning
Frequentist	confidence interval	95% certain $\beta_1 \in (51.4, 64.4)$
Bayesian	credible set	$P(56.6 \leq \beta_1 \leq 59.2) = 0.95$

- Both classic and Bayesian methods indicate feel like temperature has a significant association with number of registered riders.

Summary

- Bayesian statistics treats probability as a **degree of belief** and incorporates probabilistic reasoning into analysis.

Summary

- Bayesian statistics treats probability as a **degree of belief** and incorporates probabilistic reasoning into analysis.
- The biggest difference between frequentist and Bayesian is that frequntist views a parameter **fixed** while it's a **random variable** in Bayesian.

Summary

- Bayesian statistics treats probability as a **degree of belief** and incorporates probabilistic reasoning into analysis.
- The biggest difference between frequentist and Bayesian is that frequntist views a parameter **fixed** while it's a **random variable** in Bayesian.
- There are various choices of priors. **Domain knowledge** and **sensitivity analysis** are crucial to obtain reasonable and consistent results.

Summary

- Bayesian statistics treats probability as a **degree of belief** and incorporates probabilistic reasoning into analysis.
- The biggest difference between frequentist and Bayesian is that frequntist views a parameter **fixed** while it's a **random variable** in Bayesian.
- There are various choices of priors. **Domain knowledge** and **sensitivity analysis** are crucial to obtain reasonable and consistent results.
- R is your good friend!

Thank you!

Happy Friday:)

Q&A

References

James M. Flegal, John Hughes, Dootika Vats, and Ning Dai. (2018). mcmcse: Monte Carlo Standard Errors for MCMC. R package version 1.3-3. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, Donald B. Rubin (2013). Bayesian Data Analysis. Chapman and Hall/CRC.

Charles J. Geyer and Leif T. Johnson (2019). mcmc: Markov Chain Monte Carlo. R package version 0.9-6. https://CRAN.R-project.org/package=mcmc

Andrew D. Martin, Kevin M. Quinn, Jong Hee Park (2011). MCMCpack: Markov Chain Monte Carlo in R. Journal of Stat Software. 42(9): 1-21.

Appendix

- Data link: `https://www.macalester.edu/~dshuman1/data/155/bike_share.csv`
- Posterior derivation in linear regression example with $\mu_{\beta_k} = 0, \sigma^2_{\beta_k} = 10^2$:

$$
\begin{aligned}
P(\beta_0|\cdot) &\propto \text{likelihood} \times \text{prior} \\
&\propto \prod_{i=1}^{n} P(y_i|\beta_0) \times P(\beta_0) \\
&\propto \prod_{i=1}^{n} \exp^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2} \times \exp^{-\frac{1}{2\cdot 10^2}\beta_0^2} \\
&\propto \exp^{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{10^2}\right)\beta_0^2-\frac{2}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1 x_i)\beta_0\right]} \\
\Rightarrow \beta_0|\cdot &\sim N\left(\frac{\frac{\sum_{i=1}^{n}(y_i-\beta_1 x_i)}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{10^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{10^2}}\right).
\end{aligned}
$$

Appendix

- Following the same technique,

$$\beta_0|\cdot \sim N\left(\frac{\frac{\sum_{i=1}^{n}(y_i-\beta_0)x_i}{\sigma^2}}{\frac{n}{\sigma^2}+\frac{1}{10^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{10^2}}\right),$$

$$\sigma^2|\cdot \sim \mathrm{InvGamma}\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right).$$